VALIDITY BY DESIGN: THE PRICE OF SAFE EARLY STOPPING IN CLOSED-LOOP BATTERY TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Finding longer-lived lithium-ion batteries means cycling candidate cells until they wear out, weeks to months of channel time each. A campaign can save that time by choosing each next candidate from earlier results and stopping unpromising tests early, but a stop is irreversible and destroys the evidence that a design would have succeeded. Conformal prediction bounds the rate of such false stops in finite samples when the calibration cells are exchangeable with the cell being judged, and Bayesian optimization breaks that exchangeability. We rebuild it by design with the ε-stream, which sends each freed channel slot to a calibration cell with probability ε, drawn from a pre-registered sequence the optimizer never reads, so the conformal weights are exactly known and the certified modes carry finite-sample guarantees for the rule as deployed. We prove a ladder of such guarantees, each priced in channel time, calibration cells or delay, together with two classical floors and one impossibility, since on calibration pools holding a fixed number of cells per design those weights admit no distribution-free guarantee exact at every level. Replaying five public battery corpora, we find that successive halving, a heuristic from prior battery-testing frameworks, falsely stops as many as 96% of good cells, while our audit-only weighted rule stays within α on every seed of the unbudgeted replays; the certified modes' covered first stops reach about 1% of off-spec cells, so at these calibration sizes a certificate costs more channel time than early stopping saves.

1 INTRODUCTION

Battery development is rate-limited by evaluation. One cycling test of a candidate charging protocol or formulation occupies a physical channel for weeks to months before its cycle life is known (Severson et al., 2019; Attia et al., 2020; Zhang et al., 2026). Closed-loop platforms therefore pair a Bayesian-optimization (BO) outer loop with early prediction. Attia et al. (2020) truncated their closed-loop evaluations at 100 cycles and fed predictions to the optimizer under homoskedastic uncertainty. Later frameworks prune partial curves by successive halving (Deng et al., 2022).

Stopping a battery test, however, is not like pausing a neural-network training run. It is irreversible, and a falsely stopped cell permanently removes a candidate chemistry from the campaign along with the evidence that it would have succeeded. The early cycles of lithium-ion cells are nearly indistinguishable *in capacity* while their lifetimes differ by an order of magnitude (Severson et al., 2019), so partial-capacity rankings (the signal halving methods act on) do not predict endpoint rankings. Two cells of the MATR corpus (Severson et al., 2019) under different fast-charging protocols (`b3c7` and `b2c27` in the released audit) hold discharge capacities 0.012 mAh apart at cycle 100, inside measurement noise on a 1.1 Ah cell, yet one dies at cycle 467 and the other lasts to $1{,}834$. Any capacity-ranking rule treats them as interchangeable. On replays of MATR, ISU-ILCC (Thelen et al., 2023) and HUST (Ma et al., 2022), successive halving, which by design stops cells whose capacity at a look falls below the running median there, kills 47–96% of good cells against a 10% budget, and even margin thresholds tuned on calibration data violate the budget by 1.3–$4.7\times$ on worst seeds (Section 5). A campaign runs once, so seed-fragile tails are not deployable.

Controlling this risk inside a live optimization loop faces three coupled challenges. *(C1) Irreversibility demands conditional, finite-sample control*: a false stop cannot be undone, so the guarantee must bind the probability of killing a *good* cell, per trial, at the calibration sizes a laboratory can af-

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

ford. *(C2) Adaptive acquisition destroys exchangeability*: conformal control needs the test point exchangeable with calibration data, which the feedback loop violates; weighted extensions relax this only given a known density ratio (Tibshirani et al., 2019; Fannjiang et al., 2022), which an in-campaign acquisition does not supply for free (§6). *(C3) Calibration is scarce*: every calibration cell costs weeks of channel time, and with few known-good cells the achievable p-value granularity may not support the budget at all. To our knowledge, no battery-loop stopping rule meets all three.

We introduce a closed-loop testing framework whose certified modes carry finite-sample false-stop control for the rule as deployed, over every decision in a block-exclusive reference construction and over the campaign's first stop in the modes that can act, up to a coupling factor we bound only on front-loaded slots. In the ε-*stream*, an ε-share of channel slots loads calibration cells from a pre-registered i.i.d. sequence, firewalled from surrogate and acquisition, so the exchangeability BO destroys is rebuilt inside the campaign and the acquisition's explicit selection probabilities supply exact conformal weights (C2); good-only calibration makes the guarantee conditional (C1); a feasibility-constrained spending schedule allocates the budget across looks, and when calibration cannot support the budget the method abstains (C3).

Contributions. (i) A decision problem. Irreversible early stopping inside an adaptive battery-testing loop, hitting-time objectives, right-censoring, channel scheduling, and acquisition feedback that breaks exchangeability (Section 2). **(ii) A price theory, and an impossibility.** What is new is the firewalled ε-stream and the three theorems it earns, first-stop control on a shared pool up to a coupling factor, exactness on a front-loaded stream at a computable void risk, and a closed-form cap on reuse inflation. Every rung of the ladder states its price in channel time, calibration cells or delay, and two of those prices are classical floors instantiated here. The sharpest result is negative. The single-ratio weights a deployed campaign uses admit *no* distribution-free certificate exact at every level on stratified pools, at any replication depth, quantifying a limitation Wieczorek (2023) identified (Sections 3–4). **(iii) A benchmark and its evidence.** A reproducible replay suite over five public battery corpora (637 cells, 244 arms) and nine stopping policies, with ablations, sweeps, stress probes, a deployment-mode audit and a solid-state case study, releasing the full validity–power–resource frontier (Section 5; Table 1, Appendix C).

2 Problem Setting

Notation. A campaign optimizes over a finite design space of *arms* $\mathcal{A} = \{a_1, \dots, a_M\}$ (charging or discharging protocols, imposed cycling conditions, formulation variables), where arm a holds n_a physical replicate cells. Evaluating an arm mounts one of its cells on a testing channel; cycling produces a capacity trajectory $q(c)$, $c = 1, 2, \dots$, observed by the decision rule at *look cycles* $c_1 < \cdots < c_J$. The objective is cycle life: the cell's life L is its corpus's shipped end-of-life label (80% state of health as the BatteryLife distribution labels it, used unmodified in the primary replays; other rules: App. C), right-censored at the cell's recorded test horizon C_x; all analysis uses the truncated label $\tilde{L} \coloneqq \min(L, C_x)$. A cell is *good* if $\tilde{L} \geq L_{\mathrm{spec}}$ for a fixed specification L_{spec} (a domain constant in deployment; our primary replays instantiate it as the pool median, one value per corpus, held identical across every seed and policy), so goodness means spec attainment *demonstrated within the recorded test*. Determinacy requires only that no cell be censored *alive* before the spec ($C_x \geq L_{\mathrm{spec}}$ whenever $L > C_x$); cells that die within their record need no horizon condition. Only exceptions whose original L still reaches the spec can flip a label; a per-corpus audit of those (zero on MATR/ISU-ILCC, two marginal HUST cells counted not-observed-good) is in Appendix C. The lab operates K parallel channels; a stop or completion (at the recorded horizon, App. C) frees a channel for the next candidate (cost = busy channel-cycles and wall-clock).

The decision problem. At each look the campaign may irreversibly STOP the running cell. Stopping a cell that would miss the spec releases its remaining test time; stopping a good cell destroys the evidence that its design succeeds. Stopping is scored by bad-cell recall and released channel-cycles; the campaign objective, arm-level cycle life, is scored by observed-endpoint regret and discovery time (Section 5). We require trial-level control of the *false-stop rate* (FSR),

$$\mathrm{FSR} \;\coloneqq\; \Pr\big[\exists\, j \leq J :\ \textsc{stop at}\ c_j \;\big|\; \tilde{L} \geq L_{\mathrm{spec}}\big] \;\leq\; \alpha, \tag{1}$$

finite-sample, distribution-free in the forecaster, and for the rule as deployed inside the closed loop, where the acquisition selects arms adaptively. Under adaptive acquisition, Section 4 delivers (1)

for every decision only in a block-exclusive reference construction that cannot stop in our replays. The other certified modes bound its first-stop restriction, the probability that the campaign's first stop kills a given good cell (on a shared pool up to a coupling factor), and the weighted rule we recommend is audit-only.

3 METHOD

The method has three components: a forecaster that carries no validity burden, a good-only conformal stopping rule with feasibility-constrained α-spending, and a calibration stream whose *design* restores the exchangeability that adaptive acquisition destroys.

3.1 FORECASTER (NO VALIDITY BURDEN)

At each look c_j a pre-campaign frozen model $f_j : (x, \phi_j(q_{1:c_j})) \to (\hat{\mu}, \hat{\sigma})$ predicts $\log_{10} \tilde{L}$ from the design vector and battery-native partial-trajectory features ϕ_j (capacity levels and fade slope, Δq variance, knee onset, and the $\Delta Q_{100-10}(V)$ variance of Severson et al., 2019). Features are strictly causal: ϕ_j uses only cycles $1{:}c_j$, so the $\Delta Q_{100-10}(V)$ feature enters at looks $c_j \geq 100$ and is absent at the cycle-50 look. On HUST the feature is excluded because its 77 cells carry 77 distinct multi-stage *discharge* protocols, so discharge-curve features encode protocol identity. The forecaster is fitted on $\log_{10}$-life, so the conformal band is multiplicative in life. A single snapshot is trained *before the campaign* on a dedicated forecaster partition (never on campaign cells, so no test cell is ever scored in-sample), and is frozen throughout. The guarantees below hold for an *arbitrary* choice of f_j, so where they apply a poor forecaster costs recall, never safety.

3.2 GOOD-ONLY α-SPENDING CONFORMAL STOPPING

Let Cal_j contain only calibration cells *known good* (completed with $\tilde{L} \geq L_{\mathrm{spec}}$, or observed alive at a cycle $\geq L_{\mathrm{spec}}$). With studentized scores $s_j(x) = (\hat{\mu}_j(x) - \log_{10} \tilde{L})/\hat{\sigma}_j(x)$ and the boundary score $\hat{s}_j = (\hat{\mu}_j(x) - \log_{10} L_{\mathrm{spec}})/\hat{\sigma}_j(x)$ of the running cell, the lower-tail conformal p-value (a good-only reference sample makes this an outlier-testing construction; cf. Bates et al., 2023) is

$$p_j \;=\; \frac{w_x + \sum_{i \in \mathrm{Cal}_j} w_i \, \mathbf{1}\{s_j(x_i) \leq \hat{s}_j\}}{w_x + \sum_{i \in \mathrm{Cal}_j} w_i}, \tag{2}$$

with weights defined in Section 3.3. Good-only calibration is what upgrades the guarantee from the joint bound $\Pr[\textsc{stop} \wedge \mathrm{good}] \leq \alpha$ (all a mixed pool can give, and a factor $1/\Pr[\mathrm{good}]$ weaker as a conditional statement) to the conditional FSR of (1): conditional on goodness, the test cell's true score is exchangeable with the good-only calibration scores, and pathwise dominance ($s_j^{\mathrm{true}} \leq \hat{s}_j$ for a good cell) transfers superuniformity to the deployed statistic.

The campaign (we call this stopping rule *Rule A*) stops a cell at the first look with $p_j \leq \alpha_j$, where the spending schedule satisfies $\sum_j \alpha_j \leq \alpha$, is optimized pre-campaign on a disjoint SCHED split (scored against the offline calibration pool) to maximize expected released channel-cycles, is *frozen before the campaign*, and is feasibility-constrained: any $\alpha_j < 1/(|\mathrm{Cal}_j| + 1)$ is re-allocated, the unweighted conformal floor (Proposition 2(i)), applied to every schedule-optimizing mode as the deployed feasibility gate, computed *once, pre-campaign*, from the offline pool's per-look sizes, never from campaign data. Withholding budget from looks a discrete statistic cannot attain is classical (Tarone, 1990; exact optimal binomial spending: Silva et al., 2020); ours is the conformal instance, with the floor decided pre-campaign and abstention (not a weaker test) when no look qualifies. When no look can attain any share of the budget the method *abstains entirely*; it refuses to certify without calibration support rather than degrade its guarantee (Section 5 shows this firing at $\alpha = 0.05$ on a single-replicate pool).

3.3 THE ε-STREAM

Adaptive acquisition makes the test cell non-exchangeable with any fixed calibration set, the standard objection to conformal guarantees inside BO. We resolve it by *design*: whenever a channel frees, an independent ε-coin decides whether the slot loads a CAL cell (its arm drawn from a *pre-registered* i.i.d. sequence with probability mass function (pmf) π_{cal} fixed before the campaign, its

Algorithm 1 Deployed weighted stopping rule (one campaign)

Pre-campaign (frozen): forecaster $\hat{f}$ on the forecaster partition; spending schedule (α_j), $\sum_j \alpha_j \le \alpha$; calibration pmf π_{cal} and arm sequence $B_1, B_2, \ldots$; block maps (theorem modes).
while a channel frees **do**
  draw $u \sim \mathrm{Unif}(0,1)$ {coin always consumed}
  if $u < \varepsilon$ **and** the pre-registered stream has a next index **then**
    load next CAL cell per B_k (block mode: from its block's reserve; a void index burns the slot for $\bar{c}$ cycles); occupy $\bar{c}$ cycles; the load is *resolved* at $\mathrm{load} + \bar{c}$ regardless of goodness, and good cells then enter Cal
  else
    load TRAIN cell $\sim \pi_n$ (softmax-EI mixture; pmf recorded)
  end if
end while
for each TRAIN cell at look c_j **do**
  if any gate fires (cold start / < 5 good records; missing selection context; incomplete assigned block; arm outside calibration support; $w_x \le 0$; exhausted supply or void) **then** continue
  $p_j \leftarrow$ weighted conformal p-value (2) with $w_i = \pi_n(a_i)/\pi_{\mathrm{cal}}(a_i)$, over the mode's eligible records, shared pool: loaded before this cell and admitted by this look's decision time; block mode: the assigned block, once fully admitted
  if $p_j \le \alpha_j$ **then** STOP (frees the channel; forecast fed to the surrogate) **else** continue
end for

replicate uniform among that arm's unused calibration cells), or a TRAIN cell chosen by the acquisition. CAL cells are firewalled (they never enter the surrogate, the forecaster's training set, or the acquisition, and are never stopped), so the calibration arm density is π_{cal} by construction. Reserving a randomized share of an adaptive allocation to keep later inference valid is a known device, forced exploration; Liang & Bojinov (2024) mix a bandit with a Bernoulli design at a user-set rate for precisely this purpose. The firewall adds finite-sample *conformal* validity rather than an asymptotic confidence sequence, with pre-registered (hence exactly known) single-ratio weights. The acquisition is a softmax over expected improvement (EI) mixed with a λ-uniform floor over non-exhausted arms, so its pmf π_n at the n-th selection event is explicit and recorded at selection time; the weights in (2) are the exact single ratios

$$w_i = \pi_n(a_i)/\pi_{\mathrm{cal}}(a_i), \qquad w_x = \pi_n(a_x)/\pi_{\mathrm{cal}}(a_x). \tag{3}$$

Weights are bounded by construction ($\sup w \le 1/\min_a \pi_{\mathrm{cal}}(a)$, a pre-campaign constant since $\pi_n(a) \le 1$ always, arm exhaustion included), and we never clip on the guarantee path. Theorem modes spend the whole budget on the final look, no data-dependent re-allocation in any theorem-covered arm (Section 4). The price is explicit. An ε share of *load events* funds calibration, a larger share of channel *time* because a CAL cell holds its channel for $\bar{c}$ (charged in full, Table 11), and early decisions abstain until enough CAL cells certify good (the resulting discovery delay is in Table 11). Two calibration-cost conventions are fixed: a calibration cell holds its channel for exactly $\bar{c}$ cycles *regardless of when it dies* (value-independent occupancy is what the coupling proof requires; the full channel-time is charged to the method in every reported number), and in the offline-calibrated screening mode the calibration pool is historical data whose acquisition cost is sunk, stated wherever that mode is reported. Deployability: fixed occupancy is slot *reservation*, not literal cycling (a dead cell may be unmounted; the proofs need only value-independent timing; Algorithm 1). The softmax temperature is adaptive, not tuned; it and every other constant are disclosed in Appendix C.

3.4 THE PRACTICAL SCREENING MODE

When a historical pool of fully tested cells exists (a common industrial situation), calibration need not be purchased online. The *weighted static mode* applies the same good-only machinery with $\pi_{\mathrm{cal}}^{\mathrm{off}}$ the offline pool's arm composition (*all* cells, the $\Pr[\mathrm{good} \mid a]$ tilts cancel in the ratio) and π_n the acquisition's recorded pmf at the test cell's own selection, giving single-ratio weights $\pi_n(a)/\pi_{\mathrm{cal}}^{\mathrm{off}}(a)$: the selection-shift correction of Section 3.3 transferred to a pool whose sampling law is historical rather than designed. Proposition 1 covers its *split* variant (composition and scores from disjoint pool halves) up to a quantified composition-error factor, but only for an i.i.d. pool, which our stratified

pre-partition is not (it reserves a fixed count per arm, the deterministic-supply family Proposition 6 rules out), so the split rows are probes, not instances. The full-pool variant reuses one pool for both and remains audit-only. We recommend the mode when an offline pool exists precisely because it converts the theorems' calibration budget into sunk cost.

4 THEORY

Our guarantee is a *ladder* of precisely scoped statements. All results are finite-sample and distribution-free in the forecaster. L_{spec} is fixed before partitioning and campaign execution; the replay instantiation, the pool median (Section 2), is a permutation-invariant pool statistic, and conditioning on such a statistic preserves the exchangeability that every rank argument uses (Lemma 1). Constructions, proofs and proof outlines: Appendix A.

Assumption 1 (Replicate sampling)**.** *Conditional on its arm a, each physical replicate cell's recorded object (trajectory, endpoint) is an i.i.d. draw from a cell law P_a (we write Q_a for the induced law of the $\bar{c}$-truncated cell object), and the* CAL/TRAIN *eligibility pre-partition is value-independent (in the deployed configurations, a per-arm replicate reservation drawn from sorted cell-id lists; the proofs need only that the split does not read cell values).*

The remaining conditions are *design choices of the deployed system*, verified in code: the firewall and the pre-registered i.i.d. CAL stream with fixed occupancy $\bar{c} = L_{\mathrm{spec}}$ (Section 3.3); the look grid $L_{\mathrm{spec}} > c_J + 5$, which every theorem here consumes and Propositions 1/3 and Theorem 5 also state; a pre-campaign frozen forecaster snapshot, and a spending schedule fixed independently of the calibration values (every theorem mode spends the whole budget on the final look); explicit-pmf acquisition; and mode-specific supply semantics. In the *block-exclusive* mode each block's physical replicate supply is *reserved pre-campaign*, so a supply shortfall (a *void* index) is a function of the owning block's own draws and deterministic quotas alone, burns its channel slot for $\bar{c}$ cycles so that runtime dynamics never depend on supply, and disables only the owning cell's admission gate. In the *shared-pool* mode the first empty-supply draw triggers permanent abstention.

Theorem 1 (Exact FSR control, all decisions)**.** *Under Assumption 1 and the design conditions, with* block-exclusive calibration *(calibration stream indices are pre-assigned to disjoint blocks of size m and each campaign cell is pre-assigned a block by its load-order index, so block composition is independent of block values and no decision realized before a cell's load consults its block; each block's replicate supply is reserved pre-campaign, making supply events block-local; Appendix A), the deployed gated rule satisfies, for every loaded slot,* $\Pr\big[\exists\, j \le J : \text{STOP AT } c_j \mid \tilde{L} \ge L_{\mathrm{spec}}\big] \le \sum_j \alpha_j \le \alpha$*, with no coupling factor. The bound is argued directly in the deployed system.*

Theorem 2 (Per-cell first-stop control, shared pool)**.** *Under the same conditions without block exclusivity (the practical deployment in which all decisions share the accrued calibration pool) for the n-th loaded slot,* $\Pr[$*the campaign's first realized stop kills slot $n \mid n$ good*$] \le \alpha \Pr[G_n^\circ]/\Pr[G_n] \le \alpha/\Pr[\mathcal{C}_n \mid G_n^\circ]$*, where G_n, G_n° say slot n is good in the deployed system and in the never-stopping, never-exhausting reference H°, and $\mathcal{C}_n$ that the two load the* same cell *at slot n, which only a preceding stop or empty-supply draw can spoil. The stop event above lies in $\mathcal{C}_n$; the factor is* exactly 1 *at slots where neither spoiler can move the load, the K that front-loading mounts at T_0 (Theorem 5).*

Proposition 1 (Weighted static mode: first-stop validity up to composition error)**.** *Let the offline pool be i.i.d. with arm law q and value laws Q_a (the offline analogue of Assumption 1), historical (never loaded in-campaign), firewalled from the acquisition, with composition estimate $\hat{q}$ and good-only scores computed on* disjoint pre-registered halves*, and let the looks satisfy $L_{\mathrm{spec}} > c_J + 5$ so that every good record is eligible at every look. Then for the n-th loaded slot, with every probability below conditional on the composition half and with $\rho = \max_{a \in \mathrm{supp}(q)} \max\{q(a)/\hat{q}(a), \hat{q}(a)/q(a)\}$ ($\rho = \infty$ if $\hat{q}$ misses a q-supported arm; at our one-replicate reserve the two halves are then support-disjoint and the mode abstains outright),* $\Pr[$*the campaign's first realized stop kills it $\mid \tilde{L} \ge L_{\mathrm{spec}}$*$] \le \rho^2\big(\sum_j \alpha_j\big)\Pr[G_n^\circ]/\Pr[G_n] \le \rho^2\alpha/\Pr[\mathcal{C}_n \mid G_n^\circ]$ *(so $\rho = 1$ removes the composition price but not Theorem 2's coupling factor, whose supply spoiler cannot act here because the pool is never loaded). ρ is exactly the price of* not knowing *the calibration sampling law, which the ε-stream eliminates by construction (Appendix A). After the first stop, feedback makes the acquisition pool-dependent; that regime is audited, not certified (open conjecture, Appendix A).*

Table 1: The guarantee–power–resource map (100 seeds, canonical campaign; "—" = structural abstention, **bold** = every worst seed within budget). The second column names the nearest guarantee; no replay row instantiates a theorem (pool-median L_{spec}, finite recorded urns; Section 4). Each stronger guarantee names its price: half the pool (Prop. 1), an ε-stream (Thm. 2), a pre-campaign burn-in with T_0 delay (Thm. 5; sized per corpus, $m = 16/24/16$; also recall 0.03 on TONGJI; it reserves more calibration replicates than the online row, and at the same reserve the online mode, the power-boosted arm of App. C, reads 0.03, the base of §5's 3.7×; its worst-seed flag reflects post-first-stop reuse, outside that theorem's scope), or $m\bar{c}$ dedicated cycles per cell (Prop. 2(ii)). A row that never stops on any corpus reads "abstains" rather than **yes**, so no flag here is vacuous. At $\alpha = 0.1$ the audited weighted rule is the highest rung that fires with material recall while keeping every worst seed within budget. Thm. 3's final-look reuse wall covers rank calibration; the weighted extension is open.

Mode	Nearest guarantee	Calibration resource	Audited recall (M/I/H)	Worst seed $\leq \alpha$
ASHA (heuristic)	none	none	0.80/0.59/0.81	no
Static conformal	audit only	offline pool (sunk)	0.54/0.39/0.09	no
Weighted, full pool	audit only	offline pool (sunk)	0.05/0.22/—	**yes**
Weighted, split	probe (Prop. 1 not instantiated)	half the pool for composition	—/—/—	abstains
Online shared	Thm. 2 (first stop, coupled)	ε=0.2 stream	0.00/0.01/—	**yes**
Burn-in front-loaded	Thm. 5 (first stop, coupled)	$m\bar{c}$ pre-campaign + T_0 delay	0.01/0.11/—	no
Block-exclusive	Thm. 1 (all decisions)	$m\bar{c} = 5L_{\mathrm{spec}}$/cell (Prop. 2(ii))	—/—/—	abstains

Proof idea. Theorem 1 is argued *directly in the deployed system*: the firewall and block-local reserved supply leave a cell's excluded pairs i.i.d., so given goodness the scores are weighted-exchangeable with single-ratio weights (Tibshirani et al., 2019; Fannjiang et al., 2022). Theorem 2 instead couples that system pathwise to H°, the same run with unlimited supply and no stops; the two coincide up to the first realized stop, and off that regime they load different cells, or none, which is what the factor prices (App. A).

Proposition 2 (Prices of validity)**.** (i) Granularity floor. *Fix any look schedule with spends α_t, $\sum_t \alpha_t \leq \alpha$. A mode that stops at look t only when an* unweighted *good-only conformal p-value computed from n_t^+ good calibration records satisfies $\tilde{p}_t \leq \alpha_t$ cannot stop at t unless $(n_t^+ + 1)\,\alpha_t \geq 1$; if $n_t^+ < 1/\alpha - 1$ at every look,* no *schedule with total spend α stops on any data.* (ii) Cost of exclusivity. *Under Theorem 1's design a cell is stop-eligible only after all m loads of its private block are admitted; each occupies a channel for $\bar{c} = L_{\mathrm{spec}}$ cycles and serves no other cell, so stop-eligibility for N cells consumes at least $Nm\bar{c}$ dedicated calibration channel-cycles, m spec-horizons per cell before a single campaign decision is enabled (Appendix A).*

Part (i) is the deployed feasibility rule, set pre-campaign: an impossibility for the *unweighted* statistic, a conservative gate for weighted modes (App. A). Part (ii) prices block exclusivity.

What is intrinsic, and what is our construction's price. Two classical prices bind every rule of their stated class. When every arm mixes good and bad cells evenly, certified recall is capped by how far good and bad prefixes separate at the looks (Prop. 5, the two-point total-variation bound of testing theory; see, e.g., Lehmann & Romano, 2005), and every non-randomized rank rule pays the $1/(n+1)$ granularity floor (Prop. 7; Vovk et al., 2005). Wieczorek (2023) noted that covariate-shift results do not carry over to stratified sampling; Proposition 6 quantifies this for the deployed single-ratio weights, which admit no distribution-free certificate exact at every level on stratified pools, at any replication depth. The remaining prices are our construction's: front-loading buys first-stop control (3.7× the power-boosted arm's audited recall, 1.5× covered) for a delay T_0 and a computable void risk (Thm. 5); rank-calibration reuse is capped at $\alpha\kappa$ under a final-look schedule ($\kappa = 4.32$ at deployed constants, weighted extension open; Thm. 3); and per-stop certificate renewal is quadratically priced within the union-bound-certified private-block class (Thm. 4).

What to run, and what the certificate covers. With an offline pool whose per-look good counts clear the conformal floor ($\alpha_j \geq 1/(n_j+1)$), run the *weighted* rule and audit; where $|\mathrm{Cal}|$ also clears Learn-then-Test's (LTT) coarser floor ($n \geq 22$ at $\alpha = \delta = 0.1$), LTT reaches ≈0.6× the weighted recall, but its training-conditional guarantee needs i.i.d. calibration data, which adaptive acquisition also breaks, so here it too is audited, not certified. Buy a certificate only for the campaign's first stop, and only if Thm. 5's delay-and-void price is affordable. Here it never paid for itself. The audited rule saves at most ∼3% of channel time, of which 70–93% accrues *after* the first stop where

Table 2: Safety–power ladder at $\alpha = 0.1$, 100 seeds: FSR mean (max seed) / bad-cell recall. **Bold**: worst seed within budget. "—": no stops fired, so the bold is vacuous (abstention; for block-exclusive the admission gate never passed). LTT is scheduled against its *own* Clopper–Pearson floor $1 - \delta^{1/n}$ at miscoverage δ=0.1, 2.3× the conformal one (App. C); so scheduled it is a close competitor where it fires ($0.6\times$ our recall, zero false stops on MATR) and abstains on four of six corpora (Tables 6 and 7). Rows ascend the ladder; BOHB, EI-Hyperband, linear extrapolation, and fixed-horizon (all violating or forfeiting endpoints): Appendix C, 50 seeds.

Policy	MATR	ISU-ILCC	HUST
ASHA (halving)	0.818 (0.927) / 0.80	0.957 (1.000) / 0.59	0.475 (0.750) / 0.81
Naive $\hat{\mu} < L_{\text{spec}}$	0.475 (0.846) / 0.96	0.180 (0.500) / 0.95	0.549 (1.000) / 0.76
Naive + tuned margin (same cal. pool)	0.034 (0.262) / 0.68	0.029 (0.130) / 0.58	0.064 (0.467) / 0.26
LTT margin (CP-UCB, own floor)	**0.000 (0.000) / 0.03**	**0.003 (0.041) / 0.13**	**0.000 (0.000) / —**
Good-only conformal, unweighted (static)	0.012 (0.167) / 0.54	0.016 (0.111) / 0.39	0.015 (0.267) / 0.09
+ weighted FCS (ours, full pool)	**0.005 (0.095) / 0.05**	**0.008 (0.050) / 0.22**	**0.000 (0.000) / —**
Held-out-composition probe	**0.000 (0.000) / —**	**0.000 (0.000) / —**	**0.000 (0.000) / —**
Shared-pool online (Thm. 2)	**0.000 (0.000) / 0.00**	**0.001 (0.041) / 0.01**	**0.000 (0.000) / —**
Block-exclusive (Thm. 1)	**0.000 (0.000) / —**	**0.000 (0.000) / —**	**0.000 (0.000) / —**

no theorem applies, so even a perfect first-stop certificate would cover under 1% of it; and for the deployed weighted rule Prop. 6 rules out a certificate exact at every level. When no look clears a floor, do not stop.

Scope of the guarantees. Post-first-stop reuse is an open conjecture (App. A; $\text{FSR} \leq \alpha + \Delta$ in the spirit of Barber et al., 2023), with the record beyond it reported in full (§5). Replay draws finite recorded urns without replacement at a pool-median L_{spec}, so the theorem modes' rows are end-to-end audits (Table 1); the one campaign-independent split runs to total abstention (Table 5), and guarantees are trial-level and marginal.

5 EXPERIMENTS

Setup. The forecaster is a per-look Gaussian-process regressor (Matérn-$5/2$ + white noise) on $\log_{10}$-life; the acquisition is Section 3.3's softmax-EI mixture; an explicit pmf is a *precondition* of the weighting machinery; argmax-EI and batch qEI are out of scope. Five public battery corpora supply the arms, MATR (Severson et al., 2019; Attia et al., 2020), ISU-ILCC (Thelen et al., 2023), HUST (Ma et al., 2022), SNL (Preger et al., 2020) and TONGJI (Zhu et al., 2022), all retrieved through BATTERYLIFE (Tan et al., 2025), totaling 637 cells over 244 arms, with replication from 26 per arm down to one (breakdowns, recount convention and provenance audits in App. C). Arms are urns of real recorded cells drawn without replacement; $K = 8$ channels, $\varepsilon = 0.2$, looks $\{50, 100, 200\}$, fixed L_{spec} at the pool-median life in recorded cycles, $\alpha = 0.1$; 100 seeds for the main table and audits, 50 for the appendix baselines, α-sweeps, ablations and probes, 20 for the solid-state cohort (full list: App. C); seed-block bootstrap confidence intervals (CIs).

Main result: recall falls as the guarantee strengthens. Table 2 reports the ladder; every number names its configuration, and the ranges below are its per-seed means (the pooled primary defined below moves them by at most one point in the last digit).

Heuristics. On these three corpora ASHA (the halving mechanism of prior battery frameworks, Deng et al., 2022) falsely kills 47–96% of good cells; BOHB and EI-Hyperband behave similarly (0.45–0.95): better acquisition cannot rescue unsafe stopping; naive thresholding reaches FSR 0.18–0.55 with worst seeds up to 1.0 (linear extrapolation and fixed-horizon truncation: App. C).

Fair tuned baselines. A margin threshold tuned on the *same* good-only pool with no finite-sample correction looks plausible on the mean (0.03–0.06) but violates on its worst seed on all three datasets ($0.26/0.13/0.47$). Adding the honest correction (the Learn-then-Test (LTT) margin rule (Angelopoulos et al., 2025; Bates et al., 2021), which stops only when a Clopper–Pearson upper bound on the good-cell miss rate clears a pre-frozen budget split over looks) restores worst-seed safety at a schedule-sensitive recall (≤ 0.026 against the conformal floor its own bound cannot clear; 0.13 against its own Clopper–Pearson floor, its Table 2 row). It is a fair competitor at $\approx 0.6\times$

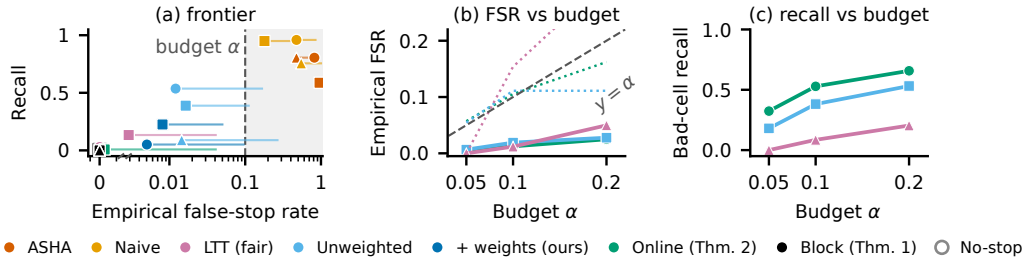

Figure 1: **(a)** Safety–power frontier, 100 seeds; **(b,c)** static mode across budgets, 50 seeds. Color = policy in (a) and dataset in (b,c); markers ● MATR, ■ ISU-ILCC, ▲ HUST; whiskers and dotted lines run to the worst seed; shading marks FSR $> \alpha$ (numbers in App. C).

the audited weighted recall, paying the binomial $\approx 2.3\times$ granularity premium (derivation and attribution: Appendix C).

Unweighted good-only conformal (static screening; shared offline pool, no selection-shift correction, outside any theorem) cuts mean FSR by an order of magnitude relative to naive thresholding (0.012–0.016; the cut versus the tuned margin is 1.8–4.4×) but still breaches on worst seeds (0.11–0.27). Acquisition cannot enter these decisions; every cell is tested and each stop is a per-cell function of the frozen calibration (verified directly in App. C), so the excursions are the finite-calibration dispersion of an at best marginally valid rule: the gap a training-conditional guarantee closes and a marginal one does not (Fig. 1(a)).

Weighted rule (full-pool feedback-covariate-shift (FCS) weights on the offline pool; held-out-composition probe: one row below): at a matched nominal $\alpha = 0.1$ this is the evaluated policy with the highest audited recall under an intact *worst-seed* budget, with max-seed FSR 0.095 (MATR) and 0.050 (ISU-ILCC) at recall 0.05 and 0.22, the 0.22 chiefly a cross-depth-of-discharge (DOD) signal (App. C); it abstains on the single-replicate HUST pool. The weights buy worst-seed stability only at a fixed nominal budget. Run at $\alpha = 0.05$ (50 seeds), the static rule reaches recall 0.32 on MATR with worst seed 0.057.

Theorem-mode audit. The online rule of Theorem 2 holds its mean at 0.000–0.001 in the canonical configuration of Table 2 (worst seed 0.041, top recall 0.008) and at 0.000–0.007 in the power-boosted one (worst seed 0.158, top recall 0.030; Appendix C). Pooling both configurations over the three corpora, the per-cell quantity Theorem 2 controls aggregates to $12/18{,}984 \approx 0.0006$ false first kills per good-cell opportunity, a figure that one first stop per campaign caps at $600/18{,}984 \approx 0.032$ for any rule; conditional on a stop occurring (which the theorem does *not* control) the first stop was false in $12/71$ stopping campaigns, and the 25 false stops after the first lie outside every theorem (Thm. 3's $\alpha\kappa = 0.43$ wall covers rank calibration only). Block exclusivity cannot stop here, since a cell's private block is admitted only after its own decisions; it is the exactness *reference* (Prop. 2(ii)).

Front-loading converts latency into power. The burn-in mode (Thm. 5, $m = 24$) reaches audited recall 0.11 on ISU-ILCC (3.7× the power-boosted online arm at that reserve) at 0.0023 false first kills per good-cell opportunity, the theorem's estimand. On the certified first stop alone recall is 0.013 against the online mode's 0.009 (other reserves and corpora: App. C). Its price is a fixed pre-campaign delay $T_0 = 2{,}163$ cycles, a stream that voids its own certificate in 36% of campaigns (Table 4), 60 post-first-stop false stops that are reuse inflation, and 5% more channel time than the replicate-matched *audited* weighted rule for 60% of its recall.

Primary metric: pooled per-trial FSR. The quantity our theorems control is the per-trial conditional false-stop probability; its estimator is the pooled rate over all good-cell trials. *Recall* throughout is bad-cell recall: the fraction of loaded bad cells the rule stops, averaged over seeds. Pooled over 100 seeds, every calibrated mode sits far inside the budget: the weighted rule at $21/3917 = 0.005$ (MATR), $45/5381 = 0.008$ (ISU-ILCC), and $0/1500$ (HUST, abstaining), descriptive pooled rates, with seed-block bootstrap intervals of the same ratio of totals ($[0.003, 0.009]$, $[0.006, 0.011]$, $[0, 0]$; trials within a campaign share calibration and scheduling), re-partition intervals over one fixed pool of recorded cells that do not model sampling of new cells. Worst-seed values are a *robustness* display, reported because a campaign runs once, and they cut both ways. Certified worst seeds reach

0.158, one weighted MATR seed hits 0.108 under the 40% budget, an exceedance the loaded-cell denominator may remove (App. C), and $14/100$ and $35/100$ campaigns (MATR/ISU-ILCC) falsely kill a good cell, so the theory delivers per-trial marginal control, not campaign-wise (App. C).

Budget tracking and what a budget buys. Across $\alpha \in \{0.05, 0.1, 0.2\}$ the static rule tracks below budget with recall monotone in α, abstaining where granularity cannot support it (Fig. 1(b,c)). Under a 40% channel-cycle budget an acquisition-matched no-stop arm isolates what stopping buys, which is very little. EI acquisition alone cuts observed-endpoint regret against uniform loading on ISU-ILCC ($40 \to 18$, paired CI $[-42, -4]$; MATR's $162 \to 123$ is within noise, $[-107, 28]$), and the static and weighted rules' seed-paired *regret* deltas against EI no-stop have CIs touching zero. Calibrated stopping does preserve the campaign's product. ASHA observes *no* top-decile endpoint in $76/100$ ISU-ILCC seeds, its observed regret exploding $18 \to 2143$ (Table 12), while the static and weighted modes match EI no-stop on discovery time and are *earlier* on ISU-ILCC (-921, -935 cycles).

Ablations (App. C). The $\Delta Q(V)$ feature carries MATR's power (static recall $0.529 \to 0.045$ without it, seed-matched) and SCHED's $+50$pp is look selection into that regime; mixed-pool calibration inflates FSR up to $2.6\times$ on HUST, so good-only calibration binds.

Generality. The ladder replicates on TONGJI (weighted recall 0.14, zero violating seeds), on the 52-cell SNL, whose gates close almost entirely, and on a real drug-permeation corpus (IVPT, recall 0.20, max seed 0.097; Tables 6 and 7).

6 Related Work

Early prediction and closed loops for batteries. Severson et al. (2019) showed early-cycle features predict cycle life; Attia et al. (2020) closed the loop with fixed-horizon truncation and Deng et al. (2022) added successive-halving pruning, neither controlling the rate at which winners are destroyed; prediction platforms sit one decision upstream (App. B). CBES (Xiao et al., 2026) calibrates confidence-bound stopping for single threshold-classification experiments (App. B).

Early stopping, and inference under adaptive collection. Successive halving, freeze-thaw and racing assume a run can be resumed or cheaply repeated; a cell cannot. Learning-curve extrapolation (Domhan et al., 2015) stops irreversibly without finite-sample control and is the lineage of our naive and extrapolation rows (survey: App. B). Risk-controlling selection (LTT/RCPS (Angelopoulos et al., 2025; Bates et al., 2021)) gives, under i.i.d. calibration data, the *training-conditional* control our marginal theorems do not. Weighted conformal prediction under covariate shift (Tibshirani et al., 2019) and feedback covariate shift (Fannjiang et al., 2022), in BO (Stanton et al., 2023), under any agent-induced shift (Prinster et al., 2024) and for threshold selection (Jin & Candès, 2026) supplies the machinery, and *our statistic is not new*. In a live campaign the exact feedback weights have factorially many terms in general (Prinster et al., 2024), and the cells a campaign completes are survivors of its own stopping rule, so they cannot calibrate it. What is new is the design that keeps those weights exactly known inside a live campaign, plus the price theory (App. B).

7 Limitations and Conclusion

Limitations. Every campaign is an offline replay, so no channel was occupied and reported power leans on the imposed cycling conditions; the same corpora served development and evaluation, the theorem-mode rows audit implementations rather than instantiate the theorems, and cross-domain evidence is one real corpus (App. C).

Conclusion. Exchangeability can be rebuilt by design, and the price is the result. At these calibration sizes it is not payable, and our rule with material recall is the one Proposition 6 denies a certificate exact at every level. What endures is the ladder (Table 1) and its floors.

Reproducibility statement

The six benchmark corpora are each provided or retrievable as follows (MATR, ISU-ILCC, HUST, SNL, TONGJI via the BatteryLife distribution (Tan et al., 2025), whose HuggingFace release is gated by an access request; IVPT ships with the code for peer review); the solid-state case-study cohort is non-public and is not required to reproduce any primary claim. Each BatteryLife input

is pinned by repository revision, pull date, and SHA-256 in a machine-checkable record shipped with the code (the pull script re-downloads at the pinned revision and fails on any hash mismatch); the shipped IVPT corpus is checksummed, and the $\Delta Q(V)$ preprocessing is validated by reproducing the strong negative feature–lifetime association reported by Severson et al. (2019), corr. -0.87 over our full 169-cell pool against their -0.93 on a log–log axis over 123 of their 124 cells. All experiments fix named seeds and rerun bit-reproducibly under the execution environment recorded in each result file (the pinned cluster image and lockfile shipped with the artifact for the campaign runs; the solid-state case study is a local run whose result file records its environment; on other hardware, results are auditable from the shipped per-seed files rather than bit-reproducible). Determinism is enforced in three layers, each of which we found necessary: (i) *code-level*, the partition draws themselves are made from sorted cell-id lists, so the assignment of cells to the four partitions does not depend on Python's per-process hash randomization, which we detected as an uncontrolled randomness source; (ii) *threading*, single-threaded BLAS (multi-threaded reductions are order-nondeterministic); (iii) *hardware dispatch*, a fixed OpenBLAS kernel target and a single OS image on the cluster, enforced by a scheduler constraint in the released launch scripts, because the cluster mixes two OS images whose libm differences flip acquisition selections, a hazard we verified directly by rerunning the full campaign on the other image (dynamic dispatch otherwise selects CPU-specific kernels whose floating-point differences perturb acquisition selections). Under this configuration two independent cluster jobs produced byte-identical result files (SHA-256 match). Several retained pre-fix campaign realizations (git-archived) preserve every ladder ordering and every budget verdict, an incidental robustness study across realizations. Per-seed arrays for every table entry are included; tables and figures are generated by scripts from result files with no hand transcription. The spending schedule, calibration stream pmf, and forecaster snapshots are frozen pre-campaign by construction in the released code, and the design assertions of Section 4 are checked at runtime wherever they apply on the cycle axis; the two relative-axis modes (IVPT's pseudo-life, the swing-relative ISU-ILCC relabeling) disable the look-ordering check, and App. C states what holds in its place. An anonymized code-and-audit artifact accompanies the submission; full code, data manifests, and exact reproduction commands will be released upon publication.

ETHICS STATEMENT

This work uses public battery cycling corpora and one non-public solid-state cohort, which is not required to reproduce any primary claim. No human subjects, personal data or animal experiments are involved. The method decides when to stop a laboratory test, and a false stop destroys the evidence about a candidate design, so we report worst-case as well as average behaviour and state where the guarantee does not hold. We see no dual-use concern beyond ordinary laboratory practice.

AI USE STATEMENT

In this work we used large language models (Claude, Anthropic) for the following tasks in the required-disclosure category of the ICLR 2027 policy: helping develop the conceptual framework (the price ladder and the validity-by-design construction), formulating and refining the mathematical claims, providing ingredients for and assisting in the writing of the proofs, proposing and refining hypotheses, designing and giving feedback on the experimental methodology, implementing the methods and the replay engine, cleaning and reformatting the public corpora into a common loader format, and interpreting results. Every corpus (the five public battery corpora, the IVPT permeation corpus and the solid-state cohort) is real recorded data with SHA-256 provenance. No cell, run or trajectory was synthesized by any means; the only transformations are the disclosed relabelings and feature constructions (equivalent-full-cycle, the IVPT hourly interpolant and pseudo-life, the solid-state life rule) and the forecaster-training noise probe of Appendix C. We did not use LLMs to generate synthetic data, for translation, or for qualitative or thematic data analysis, and the remaining required-disclosure tasks do not apply to this work. In the recommended-disclosure category we used LLMs for literature search and summarization, brainstorming, creating and editing software code and the released artifact, suggesting experimental parameters, drafting and editing manuscript sections, creating figures, formatting references, and proposing titles.

We have reviewed all AI-assisted work. Every theorem and proof was checked by the authors and additionally subjected to adversarial review by separately prompted models instructed to find errors, with each reported defect verified against the code before any change was made. Every table entry

and every quoted figure the registry lists is checked against the shipped result files by an automated registry of 394 assertions, and the shipped tables are regenerated byte-identically from those files. Every reference was verified field by field against Crossref, arXiv, or the official proceedings pages. The code carries 78 tests (76 in the artifact, which ships without the paper's figure files that the figure-clipping test reads), and the artifact's checks and tests were rerun cold on an operating system and Python version different from the results image. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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

Peter M. Attia, Aditya Grover, Norman Jin, Kristen A. Severson, Todor M. Markov, Yang-Hung Liao, Michael H. Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, Patrick K. Herring, Muratahan Aykol, Stephen J. Harris, Richard D. Braatz, Stefano Ermon, and William C. Chueh. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578:397–402, 2020.

Noor Awad, Neeratyoy Mallik, and Frank Hutter. DEHB: Evolutionary hyperband for scalable, robust and efficient hyperparameter optimization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2147–2153, 2021. doi: 10.24963/ijcai.2021/296.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Mauro Birattari, Thomas Stützle, Luís Paquete, and Klaus Varrentrapp. A racing algorithm for configuring metaheuristics. In *Genetic and Evolutionary Computation Conference (GECCO)*, pp. 11–18, 2002.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society: Series B*, 85(1):24–45, 2023.

Stefan Clarke and Bartolomeo Stellato. Conformal prediction for early stopping in mixed integer optimization. In *International Conference on Machine Learning*, 2026. arXiv:2602.01476.

Changyu Deng, Andrew Kim, and Wei Lu. A generic battery-cycling optimization framework with learned sampling and early stopping strategies. *Patterns*, 3(7):100531, 2022.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3460–3468, 2015.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning (ICML)*, pp. 1437–1446, 2018.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Alexis Geslin, Le Xu, Devi Ganapathi, Kevin Moy, William C. Chueh, and Simona Onori. Dynamic cycling enhances battery lifetime. *Nature Energy*, 10(2):172–180, 2025. doi: 10.1038/s41560-024-01675-8.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 240–248, 2016.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Ying Jin and Emmanuel J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika*, 113(1):asaf066, 2026. doi: 10.1093/biomet/asaf066. arXiv:2307.09291.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *ACM Journal of Data Science*, 2026. doi: 10.1145/3786352.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, third edition, 2005. doi: 10.1007/0-387-27605-X.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. In *Machine Learning and Systems (MLSys)*, volume 2, pp. 230–246, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Tingkai Li, Zihao Zhou, Adam Thelen, David A. Howey, and Chao Hu. Predicting battery lifetime under varying usage conditions from early aging data. *Cell Reports Physical Science*, 5(4):101891, 2024.

Biyonka Liang and Iavor Bojinov. An experimental design for anytime-valid causal inference on multi-armed bandits. *arXiv preprint arXiv:2311.05794*, 2024.

Guijun Ma, Songpei Xu, Benben Jiang, Cheng Cheng, Xin Yang, Yue Shen, Tao Yang, Yunhui Huang, Han Ding, and Ye Yuan. Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. *Energy & Environmental Science*, 15(10):4083–4094, 2022. doi: 10.1039/D2EE01676A.

Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems*, volume 6, 1993.

Yash Nair and Lucas Janson. Randomization tests for adaptively collected data. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2705554.

Yuliya Preger, Heather M. Barkholtz, Armando Fresquez, Daniel L. Campbell, Benjamin W. Juba, Jessica Román-Kustas, Summer R. Ferreira, and Babu Chalamala. Degradation of commercial lithium-ion cells as a function of chemistry and cycling conditions. *Journal of The Electrochemical Society*, 167(12):120532, 2020.

Drew Prinster, Suchi Saria, and Anqi Liu. JAWS-X: Addressing efficiency bottlenecks of conformal prediction under standard and feedback covariate shift. In *International Conference on Machine Learning*, pp. 28167–28190, 2023.

Drew Prinster, Samuel Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *International Conference on Machine Learning*, pp. 41086–41118, 2024.

Herilalaina Rakotoarison, Steven Adriaensen, Neeratyoy Mallik, Samir Garibov, Eddie Bergman, and Frank Hutter. In-context freeze-thaw Bayesian optimization for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 41982–42008, 2024.

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

Kristen A. Severson, Peter M. Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H. Chen, Muratahan Aykol, Patrick K. Herring, Dimitrios Fraggedakis, Martin Z. Bazant, Stephen J. Harris, William C. Chueh, and Richard D. Braatz. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4:383–391, 2019.

Ivair R. Silva, Martin Kulldorff, and W. Katherine Yih. Optimal alpha spending for sequential analysis with binomial data. *Journal of the Royal Statistical Society: Series B*, 82(4):1141–1164, 2020. doi: 10.1111/rssb.12379.

Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal prediction sets. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 959–986, 2023.

David Stenger, Dominik Scheurenberg, Heike Vallery, and Sebastian Trimpe. Early stopping Bayesian optimization for controller tuning. In *IEEE Conference on Decision and Control*, pp. 3747–3753, 2024.

Kevin Swersky, Jasper Snoek, and Ryan P. Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.

Ruifeng Tan, Weixiang Hong, Jiayue Tang, Xibin Lu, Ruijun Ma, Xiang Zheng, Jia Li, Jiaqiang Huang, and Tong-Yi Zhang. BatteryLife: A comprehensive dataset and benchmark for battery life prediction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5789–5800, 2025.

Robert E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522, 1990.

Muhammad Faaiz Taufiq, Jean-François Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31512–31524, 2022.

Adam Thelen, Tingkai Li, Jinqiang Liu, Chad Tischer, and Chao Hu. ISU-ILCC battery aging dataset. Iowa State University DataShare, 2023. doi:10.25380/IASTATE.22582234.V2.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022.

Jerzy Wieczorek. Design-based conformal prediction. *Survey Methodology*, 49(2), 2023.

Martin Wistuba, Arlind Kadra, and Josif Grabocka. Supervising the multi-fidelity race of hyperparameter configurations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13470–13484, 2022.

Yongrui Xiao, Mengjia Zhu, Yu Zhang, Dimitrios Tsaoulidis, and Tao Chen. Confidence-bound early stopping of experiments with sequential calibration. *Chemical Engineering Research and Design*, 230:756–766, 2026. doi: 10.1016/j.cherd.2026.05.013.

Han Zhang, Xiaofan Gui, Shun Zheng, Ziheng Lu, Yuqi Li, and Jiang Bian. BatteryML: An open-source platform for machine learning on battery degradation. In *International Conference on Learning Representations*, 2024.

Jiawei Zhang, Yifei Zhang, Baozhao Yi, Yao Ren, Qi Jiao, Hanyu Bai, Weiran Jiang, and Ziyou Song. Discovery Learning predicts battery cycle life from minimal experiments. *Nature*, 650 (8100):110–115, 2026. doi: 10.1038/s41586-025-09951-7.

Jiangong Zhu, Yixiu Wang, Yuan Huang, R. Bhushan Gopaluni, Yankai Cao, Michael Heere, Martin J. Mühlbauer, Liuda Mereacre, Haifeng Dai, Xinhua Liu, Anatoliy Senyshyn, Xuezhe Wei, Michael Knapp, and Helmut Ehrenberg. Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nature Communications*, 13:2261, 2022.

A PROOFS

Setup. Fix the campaign model of Sections 2–3: finite arm set $\mathcal{A}$; pre-campaign four-way cell partition (forecaster-training / campaign urns / calibration / schedule-tuning), the calibration split per Assumption 1; K event-driven channels; each freed slot consumes an independent ε-coin; calibration loads follow a pre-registered i.i.d. arm sequence $B_1, B_2, \cdots \sim \pi_{\mathrm{cal}}$ drawn from an independent pre-campaign stream, occupy exactly $\bar{c} = L_{\mathrm{spec}}$ cycles, and are *resolved* (admitted) at $\mathrm{load} + \bar{c}$, good cells then entering Cal. Supply semantics are mode-specific. The shared-pool mode abstains permanently from the first empty-supply draw, while the block-exclusive mode reserves each block's physical supply pre-campaign (the index-block construction below) and has no global supply event at all. The forecaster is one pre-campaign snapshot $\hat{f}$ trained on the forecaster partition only (m is reserved for calibration-block and burn-in stream sizes throughout); the schedule (α_j) is a pre-campaign constant with $\sum_j \alpha_j \le \alpha$. The looks satisfy $c_J + 5 < L_{\mathrm{spec}}$ (the form the hypotheses actually use) on every battery configuration whose results we *report*, by a margin of at least 135 cycles (tightest: the solid-state cohort, $c_J = 60$ against $L_{\mathrm{spec}} = 200$). One displayed configuration sits tighter, outside that scope: the $L_{\mathrm{spec}} = 300$ SNL spec-stress leg at 95, which we report as not run. The runtime check asserts exactly that, and is active on all but the two relative-axis modes: the swing-relative relabeling, where it holds regardless ($205 < 563$), and the inverted-domain IVPT arm, which carries the requirement on its own time axis instead, with two of 135 runs in breach (App. C). The deployed rule is *gated*: it continues (never stops) on cold start, missing selection context, an incompletely admitted assigned block, test arm outside calibration support, or nonpositive weights, and permanently once the shared pool's supply is exhausted or a burn-in void occurs.

Lemma 1 (Conditioning on permutation-invariant statistics)**.** *Let W be a random element (possibly trivial), let $Z = (Z_1, \ldots, Z_N)$ be exchangeable (or weighted-exchangeable) conditional on W, for instance independent of W, and let $T = T(Z, W)$ be permutation-invariant in Z for every value of W. Conditional on (T, W), the vector Z remains exchangeable (respectively weighted-exchangeable).*

Proof. Fix $W = \omega$ and write Z_π for the coordinates of Z permuted by π. For every permutation π and measurable sets A, B: $\Pr[Z_\pi \in A,\, T \in B \mid W = \omega] = \Pr[Z_\pi \in A,\, T(Z_\pi, \omega) \in B \mid W = \omega] = \Pr[Z \in A,\, T(Z, \omega) \in B \mid W = \omega]$, the first equality because $T(Z_\pi, \omega) = T(Z, \omega)$ pointwise and the second because $Z_\pi \overset{d}{=} Z$ given $W = \omega$. Dividing by $\Pr[T \in B \mid W = \omega]$ gives exchangeability conditional on every event $\{T \in B, W = \omega\}$ of positive probability, hence for the regular conditional distribution given (T, W) (which exists on our Polish state spaces). In the weighted case the conditional density given $W = \omega$ factorizes as $f(z) = g(z) \prod_i w_i(z_i)$ with g symmetric. Conditioning on $\{T \in B\}$ multiplies it by $\mathbf{1}\{T(z, \omega) \in B\} / \Pr[T \in B \mid W = \omega]$,

which is symmetric in z and is absorbed into g; the weight factors are untouched, so the conditional law is weighted-exchangeable with the same weights. □

The theorems are stated for a fixed, exogenous specification L_{spec}, the deployment object. The replay instantiates it as the pool median: Lemma 1 covers every direct conformal rank step exactly, *conditional* on the median (a permutation-invariant pool statistic; the values of the pool's remaining cells, which also fix the forecaster snapshot, enter as W, so each rank step holds conditional on the median and the snapshot). The ghost-value constructions (the void-index value augmentation in Theorem 1, reused verbatim by Theorem 5, and the unlimited-supply stream of H° behind Theorem 2) draw ghost values from the *unconditional* arm laws, which the median-conditioning tilts; the pool-median replay therefore does not instantiate those three proofs, and the theorem-mode replay rows are reported as an end-to-end *empirical audit* of the theorem-mode implementations, with the proven scope being fixed exogenous L_{spec}.

Lemma 2 (Value-independent selection)**.** *Under Assumption 1, any loading process that selects replicate indices by rules measurable with respect to value-independent randomness (coins, uniform-among-unused draws, other cells' bookkeeping) yields selected values i.i.d. from their arm laws given the selection pattern. This covers without-replacement urn draws. Distinct physical replicates are distinct coordinates of an i.i.d. vector, so drawing distinct indices in any value-independent order returns a fixed index subset (only without-replacement draws occur in the deployed system).*

Proof. Condition on the realized index-selection sequence; it is independent of the value vector by construction, and any fixed index subset of an i.i.d. vector is i.i.d. Exchangeability of the remaining urn under without-replacement draws is a permutation of indices, which the conditioning absorbs. □

The reference system H° (behind Theorem 2, Theorem 5 and Proposition 1; not Theorem 1, which is argued in D). H° shares every piece of randomness with the deployed system D (the same ε-coins, the same pre-registered arm sequence $B_1, B_2, \ldots$, the same replicate draws and recorded values, the same acquisition-noise draws) but *never stops any cell* (every loaded cell runs its full occupancy) and its calibration stream has *unlimited supply* by value-only ghosting at the pre-registered arms: the arm sequence $B_1, B_2, \ldots$ is defined at every index, and an index whose physical replicate supply is exhausted draws a fresh value $V_k \sim Q_{B_k}$ from its arm's extended value array, so the stream never exhausts and its pairs are i.i.d. (π_{cal}, Q) throughout (a proof device; no ghost value is generated in any reported run). (Redrawing arms at unsupplied indices would tilt the law, since supply shortfall is measurable with respect to the arm draws.) In H° no decision feedback exists, so every calibration load *eligible* at one of n's decisions is a function of coins, *other* cells' horizons, and slot arithmetic, never of the coordinates the conditioning of Lemma 3 excludes. (Eligibility binds at admission time $\leq$ decision time and $\bar{c} > c_J$, so an eligible record was loaded strictly before n; later loads may read n's own horizon and are never eligible.) Proposition 1's pool is historical and never drawn in the campaign, so there the unlimited-supply clause is vacuous and H° is simply the same run with its stops removed.

Index blocks (Theorem 1 construction, made explicit). Blocks are pre-assigned segments of the calibration *stream index*: the k-th calibration load (whatever wall-clock time it occurs at) belongs to segment $\lceil k/m \rceil$ for a pre-registered block size m, and the pre-registered cell-to-block map is *injective*: each campaign cell owns a private block that no other cell's decisions ever consult. Both assignments are deterministic functions of pre-campaign randomness and load *order*, never of any recorded value; consequently the *composition* of a block (which physical cells, which arms) is independent of that block's values, and decisions realized before a cell's load consult only other blocks. Each block's physical supply is *reserved pre-campaign*: every arm's replicates are pre-permuted and dealt round-robin to blocks, and stream index k in block b is filled from b's reserve alone, so whether index k is loadable is a function of block b's own arm draws and deterministic quotas only. An over-demanded index is *void*: at runtime it *burns* its channel slot for $\bar{c}$ cycles (channel dynamics are therefore identical on all supply paths) and it never admits, which permanently fails the owning cell's full-block-admission gate and affects nothing else. The stream ends at the deterministic cap $K_{\max} = mN$. The calibration cost of exactness is therefore explicit. Supporting N campaign cells at per-decision calibration size m requires Nm calibration loads of $\bar{c}$ cycles each (an ε-stream budget of $Nm\bar{c}$ channel-cycles) which is why the block-exclusive mode is the exactness *reference*, not the

recommended operating point. To keep the deployed and reference systems aligned at admission, the gated rule includes a *full-block-admission gate*: a cell's decisions abstain until every calibration load of its assigned block has been *resolved* (observed to $\bar{c}$, good or bad), the gate does *not* require the block to be all-good; the block's good subset supplies the calibration scores, and the good-count gate (< 5 records $\Rightarrow$ abstain) is a separate, value-dependent gate covered by Lemma 5(ii); at the deployed block size $m = 5$ the two gates together pass only on an all-good block.

Lemma 3 (Calibration-stream law in H°)**.** *In H°, fix a decision index (n, j) and let $\mathcal{G}_{n,j}$ collect all campaign randomness* except *the (a_i, V_i)* draws *(arm draws and recorded values jointly) of n's eligible calibration cells and the (a_n, V_n) draw of n itself ($\mathcal{G}_{n,j}$ retains the coins, all slot arithmetic, the pre-registered index* positions *and block maps, the arm draws of every load outside the excluded set, and the draws of every campaign cell loaded before n, those still running at the decision included, since later load times read their horizons; of cells loaded after n it retains everything but the one campaign-urn coordinate per arm the proof excludes, which nothing eligible reads. It does* not *retain the arm draws B_k at n's eligible indices (under the index-block reading, at every index of n's whole block, void ones included) nor the per-arm extension tails H°'s ghosting draws from, nor, under the index-block reading, any coordinate of the reserve arrays dealt to n's block, loaded or not, nor, under the shared-prefix reading,* any *calibration value: the proof there excludes every arm's calibration array wholesale.) The exclusion is realized on a* retained-measurable *coordinate set containing those pairs, not on the realized pairs themselves, whose location is a function of the excluded arm draws; the proof constructs it. Conditional on $\mathcal{G}_{n,j}$, the eligible calibration pairs (a_i, V_i) are i.i.d. with $a_i \sim \pi_{\mathrm{cal}}$ and $V_i \mid a_i \sim Q_{a_i}$, the test pair has $a_n \sim \pi_n$ and $V_n \mid a_n \sim Q_{a_n}$, and all pairs are mutually independent.*

Proof. Formally, realize the campaign on the product space $\Omega = \Omega_{\mathrm{coin}} \otimes \Omega_{\mathrm{sel}} \otimes \Omega_B \otimes \big(\bigotimes_a \Omega_{V,a}\big) \otimes \Omega_{\mathrm{acq}}$, ε-coins; the value-independent selection randomness of Lemma 2 (the per-arm campaign and calibration urn orders and the pre-campaign block dealing), retained in full; the pre-registered arm stream; one countably infinite i.i.d. value array per arm, whose leading coordinates are that arm's physical replicates (Assumption 1) and whose tail is the extended array H°'s ghosting reads; and acquisition noise, with the product law. In H° there are no decisions, so every load time, occupancy and admission that n's own load or its eligible records depend on is a function of coins, *other* cells' horizons, and slot arithmetic, bookkeeping that reads no excluded coordinate. Under the index-block reading three of the five excluded categories are fixed pre-campaign, before any arm is read, and the other two are located by retained randomness: the excluded arm coordinates are those at n's block's stream indices (pre-registered constants) together with the acquisition coordinate realizing a_n at slot n's load (the pmf π_n is retained, the draw from it is not) and the excluded value coordinates are, for every arm a, the whole reserve array dealt to n's block for a, every coordinate of a's extension tail, and the coordinate n's own draw would consume from a's campaign urn. The tails are excluded wholesale rather than located, because which of them a void block index reads is a function of that block's own (excluded) arm draws. Excluding all of them costs nothing in either system. In H° no decision reads a calibration value at all, and in D no deployed path reads an extension coordinate, since a void burns its slot and never admits. Both index sets are retained-measurable: the dealing and the urn order are value-independent pre-campaign randomness (Lemma 2), and slot n's load event, which locates its acquisition coordinate, is itself retained. *Which* of those coordinates a realized index reads is a function of the excluded arm draws, which is exactly what the conditioning must leave free. Under the shared-prefix reading the widening differs but the mechanism is the same. Calibration occupancy is the fixed $\bar{c}$ whatever the value, and H° never exhausts, so which indices are eligible is settled by coins, other cells' horizons and slot arithmetic. The excluded coordinates are then the arm draws B_k at those indices and the acquisition coordinate realizing a_n, together with, for every arm a, the coordinate n's own draw would consume from a's campaign urn and all coordinates of a's calibration value array, extension tail included, a retained-measurable superset of the pairs actually read, admissible here because in H° no decision reads any calibration value. The urn coordinate and the acquisition coordinate are located by retained randomness exactly as above: the urn order by Lemma 2, the acquisition coordinate by slot n's retained load event. (With exhaustion the eligible index set would itself read the B_k, several of them excluded coordinates, which is precisely what the ghosting buys.) $\mathcal{G}_{n,j}$ is then the σ-algebra generated by all coordinates of Ω except those excluded ones, and conditioning is coordinate conditioning on a product space: it leaves the excluded coordinates' product law intact (Lemma 2 absorbs the value-independent index selection). The only property used is that slot n's own load *event* (its time and the pmf π_n it draws

from, never the drawn pair (a_n, V_n), which must stay free or the lemma says nothing) and the loading, occupancy and admission of its *eligible* calibration records are measurable with respect to the retained coordinates. In the shared pool that is what the eligibility cut at $\mathrm{load} < \mathrm{load}_n$ buys. Loads after n may read V_n through n's own occupancy, and are never eligible. Block exclusivity gets the same cut for free. The full-block-admission gate demands every index of the block be admitted by $\mathrm{load}_n + c_j$, and $\bar{c} > c_J$ forces any such index to have been loaded strictly before n, so no block load's pair, load time or admission time reads V_n. (Its arms and physical replicates are in any case resolved pre-campaign from stream randomness, which is what frees the block's *law* of n.) The conclusion holds verbatim in any system where the property has been established; in D under block exclusivity it is Theorem 1's facts (a)–(b) (with (b)'s appeal to Lemma 2 for the block's reserved replicates) and load times there do depend on values through *other* cells' stops, harmlessly, because the only such stops π_n^D and n's eligible set can read precede n's load, and those cells' draws are retained. □

Lemma 4 (Per-look weighted superuniformity, good-only)**.** *In the setting of Lemma 3 (or in any system where its conclusion has been established, as Theorem 1 does inside D) restrict test and calibration cells to known-good ones. With single-ratio weights $w_i = \pi_n(a_i)/\pi_{\mathrm{cal}}(a_i)$ and $w_x = \pi_n(a_n)/\pi_{\mathrm{cal}}(a_n)$, the weighted conformal p-value* (2) *computed* in that system *on its admitted scores satisfies $\Pr[p_{n,j} \le u \mid \mathrm{good}, \mathcal{G}_{n,j}] \le u$ for all $u \in (0,1)$; we write $p^\circ_{n,j}$ for the H° instance.*

Proof. Conditioning on goodness multiplies the arm laws of both the test cell and each calibration cell by $\Pr[\mathrm{good} \mid a]$, which cancels in every density ratio; the two normalizers it leaves behind are a common constant that the scale-invariant p-value absorbs, so the good-restricted pairs are weighted-exchangeable with exactly the stated single ratios; the scores are a fixed ($\mathcal{G}_{n,j}$-measurable) function of the pairs because the forecaster snapshot is pre-campaign and frozen. The claim is then the weighted conformal quantile lemma of Tibshirani et al. (2019). Sample splitting is what keeps the ratio *single*: the acquisition reads the forecaster partition and the campaign urns but never Cal_j (the firewall), so the shift the calibration set faces is the standard covariate shift of Tibshirani et al. (2019) rather than the full feedback covariate shift of Fannjiang et al. (2022), an observation Fannjiang et al. make for weighted split conformal, and the reason a held-out calibration set is load-bearing here: were calibration data to feed the surrogate driving π_n, the reduction would fail. □

Lemma 5 (Dominance and gating)**.** *(i) For a good cell the $\bar{c}$-truncated score equals the boundary score, so in the truncated stream modes both sides of* (2) *carry the same anchor. Separately, against an* offline *pool of completed lives the deployed boundary score dominates a good cell's true score, so there the deployed statistic dominates the oracle one pathwise. (ii) Every abstention gate only removes stop events, so the gated rule's false-stop events are a subset of the ungated rule's.*

Proof. (i) is the truncation convention: for a stream cell the recorded horizon *is* the occupancy bound ($C_x = \bar{c} = L_{\mathrm{spec}}$), so good $\Leftrightarrow$ alive at $\bar{c}$, the admitted label of a censoring-certified good cell is $\tilde{L} = \min(L, \bar{c}) = L_{\mathrm{spec}}$ exactly (the boundary score the implementation computes) and the two scores coincide; offline-pool cells carry their completed lives. (ii) is immediate: a gate can only convert a stop into a continue. □

Proof of Theorem 1. We argue *directly in the deployed system D*. No alternative dynamics are coupled to it, only the static value augmentation of step (c), which leaves every deployed path unchanged. Fix a loaded slot n and define $\mathcal{G}_{n,j}$ exactly as in Lemma 3, now on D's probability space: everything except the coordinate set its index-block branch constructs, n's block's arm coordinates, for every arm that block's reserve array and extension tail, and n's acquisition coordinate and per-arm campaign-urn coordinate. Three facts hold in D under injectivity, the firewall, per-block supply reservation, and fixed calibration occupancy. *(a) π_n^D is $\mathcal{G}_{n,j}$-measurable*: the acquisition pmf at n's selection event is determined by earlier cells' values, pseudo-observations, and stop decisions, and every earlier stop decision consulted only that cell's own private block, own value, and own block's void status (injectivity plus block-local supply), all of which $\mathcal{G}_{n,j}$ retains; void indices burn their slots for the same $\bar{c}$ cycles a load would occupy, so channel timing carries no supply information; and the surrogate ingests a partial trajectory only through the forecaster pseudo-observation issued at a stop. No channel through which n's block or n's value could influence π_n^D exists. *(b)* Conditional on $\mathcal{G}_{n,j}$, the private-block pairs remain i.i.d. (π_{cal}, Q) and $a_n \sim \pi_n^D$ with $V_n \mid a_n \sim Q_{a_n}$: block *membership* and timing are functions of retained randomness (Lemma 3's argument applied in D;

other cells' stops are retained and independent of the excluded draws by (a); the reserved replicates are private physical coordinates, so Lemma 2 applies to them unchanged). *(c)* Define the full-block statistic on *all* paths by augmentation, and ghost only *values*, never arms: the pre-registered stream supplies the arm B_k at every index regardless of replicate supply, so at a void index only the value is missing. Read the extension coordinates each per-arm value array of Ω already carries ($\mathcal{G}_{n,j}$ excludes them exactly as Lemma 3 does) and let $\tilde{p}_{n,j}$ be the weighted p-value computed on n's block with any over-demanded index's value drawn from its *pre-registered* arm's extended array ($V_k^\dagger \sim Q_{B_k}$). Conditional on $\mathcal{G}_{n,j}$, the block's arms remain i.i.d. π_{cal}, voidness is a deterministic function of those arms and the quotas, and all values (physical or extension) are conditionally independent with laws Q_{B_k}, so the augmented block is exactly the i.i.d. product of (b) and $\tilde{p}_{n,j}$ is conditionally superuniform by Lemma 4. (Redrawing the *arm* at a void index would tilt the block's law, since voidness is measurable with respect to the block's own arm draws.) The deployed rule sets $p_{n,j} := 1$ whenever n's block contains a void index (there the full-block-admission gate never passes and no stop occurs) and on gate-passing paths the physical block realizes exactly its pre-registered pairs, so $p_{n,j} = \tilde{p}_{n,j}$: pathwise $p_{n,j} \geq \tilde{p}_{n,j}$, and every deployed stop event is contained in $\{\tilde{p}_{n,j} \leq \alpha_j\}$. By Lemma 4 applied conditionally (its weight function is the single ratio $\pi_n^D/\pi_{\mathrm{cal}}$, which fact (a) fixes given $\mathcal{G}_{n,j}$), $\Pr[\tilde{p}_{n,j} \leq \alpha_j \mid \mathrm{good}, \mathcal{G}_{n,j}] \leq \alpha_j$; via the containment of (c), the union bound over n's own looks and the tower property finish, with $\sum_j \alpha_j \leq \alpha$ a pre-campaign constant. The out-of-support gate needs one bookkeeping remark. Decisions fire only on $\{a_n \in \mathrm{supp}(\pi_{\mathrm{cal}})\}$; conditional on that event the test law's density ratio acquires a normalization constant that the scale-invariant p-value absorbs, and in any case the gate only removes stop events (Lemma 5(ii)), so the containment is unaffected. No stop or continue event is ever conditioned on, stops of other cells realized before n's load, the only ones π_n^D can read, are *retained* in $\mathcal{G}_{n,j}$, which is harmless precisely because they are functions of randomness independent of the excluded draws. □

Proof of Theorem 2. Up to and including the campaign's first realized stop, D coincides pathwise with H° (no prior decision has altered timing, pools, or the acquisition state); moreover, on any path where a first stop occurs, the shared-mode abstention gate ensures no empty-supply draw has yet occurred, so D's admitted calibration stream coincides with H°'s unlimited one. For the first-stopped cell n the deployed statistics therefore equal the H°-statistics at every look: $\{\text{first stop kills } n \text{ at } j\} \subseteq \{p^\circ_{n,j} \leq \alpha_j\}$, an event of the clean system. Write A for the event that the campaign's first realized stop kills slot n. Permanent abstention on exhaustion means that on A neither a stop nor an empty-supply draw precedes slot n's load, so there the two systems agree pathwise up to that load, all the argument uses, since a stream divergence at a load time $\geq$ slot n's admits only at $+\bar{c} > c_J$ later and so never enters n's eligible set. Hence $A \subseteq \mathcal{C}_n$, and since goodness is an attribute of the loaded cell, $A \cap G_n = A \cap G_n^\circ$; the marginal union bound over n's looks and Lemma 4 then give $\Pr[A \cap G_n] = \Pr[A \cap G_n^\circ] \leq \alpha \Pr[G_n^\circ]$. Dividing by $\Pr[G_n]$ gives the first bound, and $\Pr[G_n] \geq \Pr[G_n \cap \mathcal{C}_n] = \Pr[G_n^\circ \cap \mathcal{C}_n]$ gives the second. The factor is the price of the coupling. Off $\mathcal{C}_n$ the n-th load is a different physical cell in the two systems, or no cell at all, because a realized stop feeds a pseudo-observation to the acquisition and an exhausted supply draw loads a campaign cell where the reference loads a ghost. Pre-first-stop *continue* decisions never enter the conditioning, because the containment is pathwise. Shared calibration disarms neither spoiler (front-loading disarms both on a set that fires, ruling out stops before T_0 outright and making a preceding empty-supply draw harmless by *burning* its slot rather than diverting it (Theorem 5)) and App. C reports how often its campaigns realize no stop at all. □

Proof of Proposition 1. Up to and including the campaign's first realized stop the deployed system coincides pathwise with the no-stop reference run on the same realizations: no stop has yet freed a channel, injected a soft-rejection pseudo-outcome, or otherwise altered the surrogate state, and the pool's cells are never loaded, so every acquisition pmf π_n realized in this regime is a function of campaign data alone, independent of the pool by the firewall. (This is where the first-stop restriction is used. A realized stop is a function of the shared pool scores, and every later π_n inherits that dependence.) Condition on the composition half: $\hat{q}$, hence the deployed weight function $\tilde{w}(a) = \pi_n(a)/\hat{q}(a)$ at any pre-first-stop decision, is measurable with respect to campaign data and the conditioning. The score-half pairs remain i.i.d. (q, Q) and independent of both, so the p-value p computed with the *true* weights $w(a) = \pi_n(a)/q(a)$ is conditionally superuniform given goodness (Lemma 4 with the score half as the calibration set and q in place of π_{cal}; the pool's scores are its completed lives, which Lemma 5(i) covers). Pathwise, each deployed weight satisfies

$\tilde{w}(a) \in [w(a)/\rho,\ \rho\, w(a)]$, so with $\tilde{p}$ the deployed p-value, numerator and denominator distort by at most ρ each: $\tilde{p} \geq p/\rho^2$. Hence $\Pr[\tilde{p} \leq \alpha_j \mid \text{good}] \leq \Pr[p \leq \rho^2 \alpha_j \mid \text{good}] \leq \rho^2 \alpha_j$; the first-stop event at look j is contained in $\{\tilde{p}_j \leq \alpha_j\}$ (Lemma 5(ii)). The deployed statistic uses the boundary score, which dominates a good cell's true score (Lemma 5(i)), so the substitution is conservative. The deployed arm does not use the re-spreading gate of Proposition 2 at all. It runs a separate rule that places the whole budget on the *latest* granularity-feasible look, and that rule reads the score-half counts, so α_j is not *a priori* constant given the conditioning. The coverage hypothesis is what removes that dependence. With $L_{\text{spec}} > c_J + 5$ every good record is eligible at every look, so the count is equal across j, the rule returns either $(0, \ldots, 0, \alpha)$ or all-zero, and the stop event is contained in $\{\tilde{p}_J \leq \alpha\}$ with the *constant* α. (Without it the eligible half is the good-*and*-long-horizon subpopulation, whose value law differs from the test cell's and breaks Lemma 4 at every ρ.) The union bound over the cell's looks finishes. ρ is measurable with respect to the conditioning, so the bound is conditionally valid as stated. Finally, superuniformity is established in the reference run, so the transfer to the deployed conditioning carries $\Pr[G_n^\circ]/\Pr[G_n]$ exactly as in Theorem 2's proof. Every probability in that transfer is taken under the composition-half conditioning, the two goodness probabilities included. The half enters the deployed weights and hence which cell slot n is, so it may not be dropped from $\Pr[G_n]$, and keeping it on both makes the cancellation to $1/\Pr[\mathcal{C}_n \mid G_n^\circ]$ exact with no independence claim needed. Of $\mathcal{C}_n$'s two spoilers only a preceding stop can act here (through the acquisition feedback it injects) since a pool that is never loaded in the campaign has no supply draw to exhaust. □

Proof of Proposition 2. (i) The unweighted good-only conformal p-value computed from n_t^+ calibration records takes values in $\{k/(n_t^+ + 1) : k = 1, \ldots, n_t^+ + 1\}$, so $\tilde{p}_t \geq 1/(n_t^+ + 1)$ deterministically, and $\{\tilde{p}_t \leq \alpha_t\}$ is empty unless $(n_t^+ + 1)\,\alpha_t \geq 1$. If $n_t^+ < 1/\alpha - 1$ at every look then $n_t^+ + 1 < 1/\alpha$, so for any schedule with $\alpha_t \leq \sum_s \alpha_s \leq \alpha$ we get $(n_t^+ + 1)\,\alpha_t < \alpha_t/\alpha \leq 1$: every look's stop event is empty. The statement is an impossibility for this *unweighted* statistic only. The deployed weighted p-value of (2) attains $w_x/(w_x + \sum_i w_i)$, which is not $1/(n_t^+ + 1)$ and not distribution-free, so the weighted modes run the same floor as a *conservative design gate*. The gate returns one of three schedules (a proportional re-spread over the surviving looks, the whole budget on the largest-n_t^+ look (earliest on ties), or all-zero) each with $\sum_j \alpha_j \leq \alpha$ pathwise. None is a constant. All three read the offline pool's per-look good counts n_t^+. What the union bounds of Theorems 1/2 need is that α_j be fixed given the conditioning. The theorem-covered arms meet this without the gate: they spend the whole budget on the final look, a pre-campaign constant, so their proofs never condition on the pattern that n_t^+ reads. The SCHED optimizer, which scores look-subsets against the calibration *values*, is run by no theorem-covered arm for the same reason. The all-zero case only removes stop events (Lemma 5(ii)). *(ii)* Stop-eligibility of cell i requires every index of its private block admitted (the full-block-admission gate), i.e., m calibration loads each observed for $\bar{c}$ cycles. Blocks are disjoint by injectivity of the index assignment and reserved supply is never shared, so the per-cell costs add: at least $Nm\bar{c}$ calibration channel-cycles for N stop-eligible cells, with $\bar{c} = L_{\text{spec}}$ by the occupancy convention of Section 3.3. □

Proposition 3 (Exact arm-matched validity under the stratified offline pool)**.** *Under Assumption 1, let the offline pool reserve exactly $c(a)$ replicate cells per calibration arm a (the deployed pre-partition), firewalled from the campaign, with the forecaster snapshot frozen pre-campaign and the schedule (α_j) fixed independently of the pool's and the campaign's values (the superuniformity below is marginal over the pool, so a schedule tuned against these calibration values voids the union bound over looks), and let goodness imply look-coverage ($L_{\text{spec}} > c_J + 5$, the 5 being the implementation's five-cycle trajectory margin beyond the final look; satisfied on every battery configuration we report).* IVPT *is outside this form. Its spec lives on a pseudo-life axis while its looks index hours, so the inequality does not apply there and the arm is treated separately (App. C). For campaign cell n with recorded arm a_n, let $\text{Cal}_j^{a_n}$ be the known-good calibration records of arm a_n at look j and*

$$p_{n,j}^{\text{arm}} = \frac{1 + \#\{i \in \text{Cal}_j^{a_n} : s_j(x_i) \leq \hat{s}_j\}}{|\text{Cal}_j^{a_n}| + 1},$$

gated to continue whenever $\text{Cal}_j^{a_n} = \emptyset$. Then, for every fixed *campaign cell n, conditional on the snapshot and $\{\tilde{L}_n \geq L_{\text{spec}}\}$, $\Pr[p_{n,j}^{\text{arm}} \leq t] \leq t$ for every look and all $t \in (0, 1)$, and stopping on*

$p^{\mathrm{arm}}_{n,j} \le \alpha_j$ *gives* $\Pr[\textsc{stop cell}\ n \mid n\ \textit{good}] \le \alpha$ *(exactly, with no weights, no composition estimate, and no ρ; marginal per fixed cell, never conditional on the campaign history (as in Theorem 2), though inventory-indexed rather than load-indexed) and, because the frozen pool is never adaptively reused, covering* every *stop of cell n, not only the campaign's first (Remark below). Moreover this is the* only *p-value the weighted-exchangeability calculus licenses for this design: with deterministic slot arms and the arm observed, each permutation's likelihood in the weights of Tibshirani et al. (2019) factors into the product of every cell's own arm density, which does not depend on the permutation and cancels, times an indicator that the permutation preserves slot arms, so the weights reduce to $\mathbf{1}\{a_i = a_n\}$ under every law. That is the arm-matched (Mondrian) p-value above (Vovk et al., 2005); the full-pool single-ratio weights of* (2) *are not superuniform here (Example below).*

Proof outline. Fix cell n: its recorded arm a_n and value V_n are attributes of the cell's own draw, independent of the pool. The statistic $p^{\mathrm{arm}}_{n,j}$ is a deterministic function of (V_n, the pool, the frozen snapshot) (defined pathwise whether or not the campaign ever loads n) and every deployed stop of n at look j is contained in $\{p^{\mathrm{arm}}_{n,j} \le \alpha_j\}$, so no conditioning on the campaign history is needed. Condition on the snapshot and the pool's goodness-and-coverage pattern G: given G the surviving slot values are independent with laws $Q^+_{a_i}$ (coordinate conditioning on the product space of Lemma 3's proof; Lemma 2 absorbs the value-independent reservation, not Lemma 3's conclusion, whose i.i.d. arm law is exactly what deterministic slot counts deny). Given further $a_n = a$ with $\mathrm{Cal}^a_j \neq \emptyset$ (a (G, a_n)-measurable gate) the test's true score and the same-arm good scores are i.i.d. images of Q^+_a under the frozen score map: the $\Pr[\mathrm{good} \mid a]$ tilt is common within the arm, trivializing Lemma 4's cancellation. Plain exchangeability gives superuniformity; the boundary score substitutes conservatively (Lemma 5(i)); gates only remove stops (Lemma 5(ii)); the tower property over (G, a_n) and the union bound over the frozen schedule finish as in Theorem 1. Feasibility is Proposition 2(i) *per arm*: $p^{\mathrm{arm}} \ge 1/(c(a)+1)$, so at the deployed $c(a) \le 7$ and $\alpha \le 0.1$ the mode abstains structurally (at $\alpha = 0.2$ the deeper reserves clear the floor). □

Remark (why marginal, not history-conditional). Conditioning on the campaign history is not available when calibration records are reused. Past stops are functions of the same records and tilt their conditional law. A one-arm miniature makes this exact: with M the minimum of nine good calibration scores and i.i.d. uniform test scores, $\Pr[X_1 < M] = 1/10$, yet $\Pr[X_2 < M \mid X_1 < M] = \mathbb{E}[M^2]/\mathbb{E}[M] = 2/11 > 1/10$. Exactly this inflation is the open conjecture's regime; the conditional-on-stop record is reported, not certified (Section 5). *Fixed* means inventory-indexed *in this proposition*: n names a physical cell of the pre-partitioned campaign inventory, whose arm is its own attribute, and the bound is proved pathwise, never conditioning on n being loaded. The weighted, load-indexed results (Prop. 1, Thms. 1/2/5) read n as the n-th *loaded* slot, whose arm is drawn from π_n, which is what Lemma 4 needs, and without which they are false: at $\pi_{\mathrm{cal}} = \pi_n = (\frac{1}{2}, \frac{1}{2})$, $m = 9$ and a misranking forecaster, an inventory-fixed arm-1 cell stops with probability $\mathbb{E}[1/(N_1+1)] = 1023/5120 = 0.20$, twice α, while the loaded slot stops with probability exactly α. The n-th *loaded* cell is a different estimand and no claim is made for it *by this proposition*, whose shared frozen pool cannot support it, Theorems 1/2 do certify a load-indexed slot, which is precisely what block privacy and first-stop containment buy. That loaded slot's arm is chosen by a history that reads past stops, and such routing inflates its stop rate, two arms with one calibration score each, one look at $\alpha = \frac{1}{2}$, load arm 1 first and rechoose arm 1 iff it stopped: the second slot stops with probability $\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{12}$, while every *fixed* cell in that construction still stops with probability at most $\frac{1}{2}$.

Theorem 3 (Reuse inflation is capped)**.** *Assume Assumption 1 and the pre-registered stream designs of Theorems 2/5, admitted good-cell scores conditionally i.i.d. given the goodness pattern Γ (of the calibration stream* and *the campaign inventory, both realized pre-campaign in the replay), the firewalled test cell drawn from the same good-conditional law, continuous score laws, with a frozen schedule (α_j) and the deployed cold-start gate a_0 (no decision fires with fewer than a_0 admitted* good *records). Then for every fixed campaign cell n,*

$$\Pr[\, n \text{ ever } \textit{stopped} \mid n\ \textit{good},\ \Gamma\,] \ \le\ \sum_j \alpha_j\, \kappa\big(\alpha_j, K_0^{(j)}\big), \qquad K_0^{(j)} = \max\{a_0,\ \lceil 1/\alpha_j \rceil - 1\},$$

$$\kappa(\alpha, K_0) = 1 + \alpha^{-1} \int_\alpha^1 \min\Big(1,\ \tfrac{e^{-K_0\,\mathrm{kl}(\alpha,s)}}{1-e^{-\mathrm{kl}(\alpha,s)}}\Big)\, ds, \quad \mathrm{kl}(\alpha, s) = \alpha \ln \tfrac{\alpha}{s} + (1-\alpha) \ln \tfrac{1-\alpha}{1-s},$$

uniformly over pool growth and over arbitrary (pool-reading, adversarial) loading and decision timing. The bound is marginal per fixed cell and covers all *stops, not only the campaign's first, so linearity gives* $\mathbb{E}[\#\textit{false stops} \mid \Gamma] \le N_{\text{good}}\,\alpha\kappa$ *under the final-look schedule. At the deployed constants (*$\alpha = 0.1$*,* $a_0 = 5$*):* $\kappa = 4.32$*, an all-stops wall of* 0.43 *per good cell. Because* $K_0 = \max\{a_0, \lceil 1/\alpha \rceil - 1\} = 9$ *here, the deployed gate* $a_0 = 5$ *is inactive and* κ *is* $\kappa(\alpha, 9)$*; the gate becomes the reuse-inflation knob only once it exceeds the granularity floor (*3.32 *at* $a_0 = 20$*,* 2.50 *at* $a_0 = 50$*). Fixed offline pools need no wall for rank-type p-values. A prefix length* K *(the count of admitted good records, not the channel count) that is constant gives* $\Pr \le \alpha$ *outright (Proposition 4's containment).* Scope*: the theorem covers* rank-type *p-values, plain or arm-matched ranks whose compared score is exchangeable with the prefix. The deployed single-ratio* weighted *p-value is not covered. Under adaptive* $\pi_n \neq \pi_{\text{cal}}$ *the cross-arm law of the pooled prefix differs from the test cell's, the binomial reduction fails, and the weighted extension (a conditional-coupling lemma) is open. We do not apply this wall to the weighted experimental rows.*

Proof outline. The stream loads in pre-registered order, so the calibration set at any decision is a *prefix* of one fixed sequence: history shifts *when* admissions happen, never which or in what order, and adaptivity's only power over the p-value is the prefix length $K \ge a_0$. A deployed stop of n at look j requires $p^{(K)}(\hat{s}) \le \alpha_j$ at the realized K, so $\{\text{stop at } j\} \subseteq \bigcup_{K \ge K_0^{(j)}} \{p^{(K)}(\hat{s}) \le \alpha_j\}$, a pathwise containment, valid for every randomized, score-observing timing strategy, with K_0 collecting the gate and the granularity no-fire floor ($1 + N_K \le \alpha(K+1) \Rightarrow K \ge \lceil 1/\alpha \rceil - 1$). Under these stream designs an admitted good record carries $\tilde{L} = \bar{c} = L_{\text{spec}}$ exactly, so its score *is* the same L_{spec}-anchored boundary score the test cell contributes (Lemma 5(i)); the same-law hypothesis then makes $s = F(\hat{s})$ uniform, and no residual-score substitution is needed (that substitution would change the reference law, not relax the event). Conditional on the goodness pattern and on s, the prefix counts N_K are $\text{Bin}(K, s)$ (conditional i.i.d.; probability-integral transform), and $p^{(K)}(\hat{s}) \le \alpha$ forces $N_K < \alpha K$; the Chernoff exponent at that threshold is exactly $\text{kl}(\alpha, s)$, a geometric sum over $K \ge K_0$ bounds the union in K, the region $s \le \alpha$ contributes at most α, and integrating over s assembles $\alpha\kappa$. A union over looks finishes. □

Remark (what the wall does and does not bound). The wall is a statement about a *marginal per-cell* quantity: for a fixed campaign cell, the probability of ever being stopped given it is good, hence $\mathbb{E}[\#\text{false stops} \mid \Gamma] \le N_{\text{good}}\,\alpha\kappa$ under the final-look schedule (κ rises as the budget is split: $\sum_j \alpha_j \kappa_j$ is 0.43 on one look but 0.52 on two equal ones). It bounds no realized per-seed ratio (Lemma 6 shows those are not controlled) so the audited figures here are reported as *position*, not as a test of the bound. Over the weighted burn-in arms pooled all-stop rates reach 0.025 with worst-seed ratios 0.19; over the plain-rank composition-proportional arms, family maxima of 0.075 and 0.4286 (attained on different rows). Neither family is a theorem instance. Both acquire with $\pi_n \neq \pi_{\text{cal}}$ from the first load (App. C), the condition the Scope above excludes. We make no claim about the attainable maximum; what the theorem certifies is that inflation *saturates* in K rather than growing. The proof uses the stream's conditional i.i.d. structure and we do not know whether a uniform-in-growth cap survives under exchangeability alone. The conjecture (exactness after the first stop) remains open; what is closed is the possibility that reuse inflation grows without bound.

Proposition 4 (Per-cell validity is cheap on a fixed pool)**.** *For rank-type p-values on a fixed, front-loaded good pool of size* m *with final-look schedules, every fixed campaign cell* n whose good-conditional score is jointly exchangeable with the pool's good scores *(same law* and *the firewall's independence, the hypothesis of Theorem 4 below; equal marginals alone admit dependent couplings that break the bound, and a test arm with a different law breaks it at any budget) satisfies* $\Pr[n \textit{ ever stopped} \mid n \textit{ good}] \le \lfloor \alpha(m+1) \rfloor / (m+1) \le \alpha$*, unlimited stops, unbounded reuse, supply* $1/\alpha - 1$*. The stop event is pathwise contained in* $\{n$*'s score falls below the pool's* $\lfloor \alpha(m+1) \rfloor$*-th smallest good score*$\}$ *(the rank-type p-value* $(1 + \#\{s_i \le \hat{s}\})/(m+1)$ *is lower-tail), one fixed event independent of all selection (the firewall), so reuse costs* nothing *at this semantics. What reuse loses is the per-epoch* renewal *of first-stop-type exactness (later stops are conditionally inflated (the* $2/11$ *Remark)), and pricing that loss is the next theorem's job. (The weighted deployed mode is excluded. After a stop,* π_n *is pool-dependent, which is exactly why Theorem 5 stays first-stop-scoped.)*

Lemma 6 (Equal-urn mounted bridge)**.** *Let* A^* *be a declared arm set with* π_{cal} *exactly uniform on* A^* *and* exactly one *test cell per declared arm; let the phase-0 stream draw* m *arms i.i.d.* π_{cal} *with replacement (any over-demand of an arm's* r*-cell calibration supply voids the campaign), and*

let the test sequence $B_1, \ldots, B_{|A^*|}$ be i.i.d. π_{cal} from an independent pre-campaign RNG, slot k mounting *arm B_k's cell iff $B_k \notin \{B_1, \ldots, B_{k-1}\}$ (else a pre-resolved skip); decisions use the plain pooled rank on the stream's admitted good records with the whole budget on the final look and a cold-start gate. Assume: (H1) cells within an arm are i.i.d. given the arm, independent across arms; (H2) both RNG sequences are independent of all values and never read urn occupancy; (H3) equal urns and uniform π_{cal} as above; (H4) L_{spec} is a fixed campaign-independent constant; (H5) $L_{\mathrm{spec}} > c_J + 5$. Then for every slot k, marginally over the slot's draw and both void events, $\Pr[$slot k's cell ever stopped $\mid$ mounted, good$] \le \alpha$ (the per-cell estimand of Section 2 restricted to tested cells) and pooled counts obey $\mathbb{E}[\#\text{false stops}] \le \alpha\, \mathbb{E}[\#\text{good mounted}]$. Per-seed FSR ratios and their seed means are* not *controlled; nor are arm-conditional or realized-stop-conditional rates; unequal urns or non-uniform π_{cal} void the lemma (two arms of capacity $(7, 1)$, 8 mounting draws against those urns and $|\mathrm{Cal}| = 8$, reach 0.52 against $\alpha = 1/3$; the exact enumeration ships with the code).*

Proof sketch. $\mathrm{mounted}_k$ is a function of the arm draws alone, invariant under arm relabeling, and the draw vector's law is relabeling-invariant (H2, H3), so $\Pr[B_k = a \mid \mathrm{mounted}_k] = 1/|A^*|$. Conditioning on goodness tilts this law by the per-arm good rates, the same tilt the stream's admitted good records carry, so both sides draw from one good-conditional arm mixture (H1). Give every voided stream occurrence a phantom value drawn i.i.d. from its arm's law (a coupling, by H1/H2): in the extended system the goodness pattern has *independent* coordinates (conditioning on the realized {no void} instead would couple arms through the supply cap; with two arms of six reserved cells each, uniform π_{cal}, twelve stream slots, good rates $(1, \frac{1}{2})$, arm-separated good scores (every arm-1 score above every arm-2 score) and $\alpha = 0.15$, no void forces six slots per arm, so on $\{n_g = 6\}$ the pool holds arm-1 scores only while the mounted good cell comes from arm 2 with probability $\frac{1}{3}$ and then stops surely, giving $\frac{1}{3} + \frac{2}{3} \cdot \frac{1}{7} = \frac{3}{7}$ against a $\frac{1}{7}$ rank bound), and given the extended pattern the good scores plus the mounted-good slot score are exchangeable. Every stop is pathwise contained in {no void}$\cap${extended final-look rank event} (zero-budget looks cannot fire; gates only remove stops), whose probability is at most α by the rank bound averaged over extended patterns; intersection only shrinks it. Reuse is free at this semantics. All of a mounted cell's stops sit in one fixed event. □

Theorem 4 (Epoch-reset: per-stop certificate renewal, quadratically priced)**.** *Pre-register E disjoint consecutive blocks of the phase-0 stream; epoch e computes rank-type p-values on block e's good records only, at budget α_e on the final look; the first realized stop of an epoch permanently retires its block; a void index kills its own block; with no live block the rule abstains. Assume each fixed test cell's good-conditional score is* exchangeable with every block's good scores *(same good-conditional law, independence alone does not suffice: with a test arm whose law differs from the blocks', independent pooled ranks fail at levels of $1 - 0.9^{10} \approx 0.65$ and worse for $m = 9$, $\alpha = 0.1$; witness: test law $\mathrm{Unif}(0, 0.1)$, and a lower point mass drives the level towards 1), and the firewall seals admissions (the acquisition never reads calibration). Then:* (i) *every realized stop is some epoch's first realized stop, and the certificate it consumes is the* fixed-cell *one: for every fixed cell n and epoch e, $\Pr[n$ stopped in epoch $e \mid n$ good$] \le \alpha_e$, hence $\Pr[n$ ever stopped $\mid n$ good$] \le \sum_e \alpha_e$: the epoch-e stop of n is pathwise contained in $\{n$'s score falls below block e's $\lfloor \alpha_e(m_e{+}1) \rfloor$-th smallest good score, m_e its admitted good count$\}$, the firewall makes that score independent of every block, and the union over epochs is immune to adversarial timing (same-tick races are serialized by the replay's deterministic event order). The control is marginal per cell; the conditional probability that a* realized *stop is false is not controlled, with one epoch, $m = 9$, and an all-good stream every realized stop is false by construction, and some cell stops with probability $N/(N{+}9)$ over N cells while every fixed cell's risk stays at 0.1.* (ii) *Within the class of pre-registered private-block schemes certified by per-epoch exactness plus the union bound (non-randomized rank rules, one certified stop per block) covering S stops requires total good reserves $\ge S^2/\alpha - S$ (per-epoch granularity floor $1/\alpha_e - 1$ and Cauchy–Schwarz over $\sum_e \alpha_e \le \alpha$), and the equal-split construction attains it. The class restriction is load-bearing. At per-cell semantics alone, $\Theta(S \log S/\alpha)$ private-block schemes exist (certified through $\mathbb{E}[\max_e V_e] \le \alpha$, V_e the blocks' exceedance variables), and Proposition 4's reused pool costs $\Theta(1/\alpha)$; the quadratic price is the price of renewing the fixed-cell certificate at every stop, not of validity.*

Proof outline. (i) Blocks are pre-registered and disjoint, and retirement is triggered only by a realized stop, so at any decision exactly one block is live and each block issues at most one certified

stop. Every realized stop is its epoch's first. Fix a cell n and an epoch e. Its epoch-e stop event is contained in $\{n$'s score falls below block e's $\lfloor \alpha_e(m_e{+}1) \rfloor$-th smallest good score$\}$, the gates only remove stops (Lemma 5(ii)) and the whole epoch budget sits on the final look. Under the same-law hypothesis n's score is exchangeable with block e's good scores, and the firewall makes it independent of the block, so that event has probability at most α_e; the union over epochs gives $\sum_e \alpha_e$. Timing cannot help an adversary. The containing events are fixed per block and the replay serializes same-tick races. *(ii)* In the stated class each certified block must clear the per-epoch granularity floor, i.e. hold at least $1/\alpha_e - 1$ good records, so covering S stops costs $\sum_{e \le S}(1/\alpha_e - 1)$. Minimizing subject to $\sum_e \alpha_e \le \alpha$, Cauchy–Schwarz gives $\sum_e 1/\alpha_e \ge S^2 / \sum_e \alpha_e \ge S^2/\alpha$, hence total reserves $\ge S^2/\alpha - S$, attained by the equal split $\alpha_e = \alpha/S$. Outside the class the bound fails. Certifying through $\mathbb{E}[\max_e V_e] \le \alpha$ rather than the union bound admits $\Theta(S \log S/\alpha)$ schemes (the maximum of S exceedance variables concentrates like $H_S/(m{+}1)$, with H_S the S-th harmonic number), and Proposition 4's single reused pool costs $\Theta(1/\alpha)$. □

The practical wall (pilot). The pilot runs the *deployed weighted* epoch variant, an engineering probe of Theorem 4's construction, not a literal instance (the weighted coupling lemma is open). On the richest feasible configuration (ISU-ILCC, three reserved replicates per arm, 100 seeds) it is *dominated*: $E{=}2$ blocks of $m{=}40$ record $16/3$ true/false epoch-first kills at recall 0.009 (max seed 0.056), $E{=}3$ records $5/0$ at 0.003, while the single-epoch burn-in at the same reserve records $48/8$ at 0.026, the quadratic supply demand collides with the void tail (second-block survival 0.035 at the granularity-feasible size) and buying more covered stops yields fewer. On this evidence sequential coverage is unaffordable on our corpora for the deployed variant; the practical frontier stays with first-stop certification plus audit (and, for rank calibration, the reuse wall). Two by-products stand: at the richer reserve the single-epoch covered-regime audit point reads 0.026 versus 0.013 at two replicates (a denominator effect (48 versus 47 covered kills against a bad-cell universe roughly halved), not more certified stops) and TONGJI's covered-regime audit reads 0.012 ($26/0$ at $m{=}12$; pool-median spec) with the frozen threshold 350 reading 0.014 ($29/0$, mean FSR 0.004; Table 4, †), a frozen-threshold replay point on a second corpus; the rule reads the same pool's endpoints, so it fixes the constant without solving specification selection.

The composition-proportional audit (support-renormalized). Table 3. Sampling the acquisition from π_{cal} itself would make every single-ratio weight 1 and put the plain pooled rank inside Theorems 3/5 and Proposition 4; the deployed partitions refuse the premise. At three replicates the test pool covers 37 of ISU-ILCC's 61 calibration-supported arms in every seed (46–54 at two; 20–30 of 34 on MATR; 6–10 of 10 on TONGJI, complete at the first load in $4/100$ seeds, none of which fire), so every stop reported in Table 3 occurred with $\pi_n \neq \pi_{\mathrm{cal}}$ from its campaign's first load, and the plain pooled rank is not the covered statistic. Every row is an *audit*, not a theorem instance. The audit shows the unweighted rule under composition-proportional acquisition firing hard (recall 0.52 on ISU-ILCC, 0.044 true first stops per bad cell at the frozen threshold 700) with every mean per-seed FSR inside α but the per-seed ratio dispersing (worst seeds 0.23–0.43, up to $33/100$ violating α), one stream draw coupling every decision of a seed. A *paired* arm reruns the same acquisition with the deployed weighted single-ratio decision, same RNG, same ordinal loaded-cell sequence (trace-audited; stopping shifts later load *times*, not the order): worst seed 0.059, zero violating seeds, recall 0.02–0.03 (Table 3). Holding the selection path fixed, the weights concentrate the per-seed ratio and price the concentration in power; the dispersion difference is attributable to the decision statistic. TONGJI's thin pooled supply all but closes the gate. It fires in $4/100$ seeds, at recall 0.01 with every first stop a true kill and no violating seed, and carries no paired weighted row. Its frozen-threshold first-stop audit runs through the weighted rule instead (the statistic of Thm. 5 at a same-pool frozen constant): 0.014 with every first stop a true kill.

Theorem 5 (Front-loaded calibration: per-cell first-stop control from a pre-campaign stream)**.** *Run the campaign in two phases.* Phase 0*: a pre-registered arm sequence $B_1, \ldots, B_m$ drawn i.i.d. from the designed pmf π_{cal} by a dedicated pre-campaign RNG is loaded on the K channels, with replicate assignment resolved pre-campaign by value-independent randomness; every stream index (including an over-demanded (*void*) index) occupies its channel for exactly $\bar{c} = L_{\mathrm{spec}}$ cycles, so phase 1 begins at the deterministic $T_0 = \lceil m/K \rceil \bar{c}$; cells observed alive at $\bar{c}$ enter the calibration set, frozen thereafter; if any index was void the certifier abstains for the entire campaign.* Phase 1*: the deployed gated rule with the frozen snapshot and schedule (no data-dependent re-allocation), single-ratio weights $\pi_n(a)/\pi_{\mathrm{cal}}(a)$ evaluated at each test cell's own loading-time pmf, and the good-only p-value* (2)*. Assume Assumption 1 and $L_{\mathrm{spec}} > c_J + 5$. Then for the n-th loaded slot of phase*

Table 3: The composition-proportional *audit* arm (100 seeds): acquisition samples π_{cal} restricted to arms with live test cells, renormalized, and stops on the plain pooled rank. This is *not* a theorem instance. At three replicates the test pool covers 37 of ISU-ILCC's 61 calibration-supported arms in every seed (46–54 at two; 20–30 of 34 on MATR; 6–10 of 10 on TONGJI, complete at the first load in only $4/100$ seeds, a set disjoint from the $4/100$ that fire), so every stop reported here occurred with $\pi_n \neq \pi_{\mathrm{cal}}$ from its campaign's first load, the single-ratio weights are not 1, and the plain rank is therefore neither the rank-type statistic Theorem 3 and Proposition 4 cover (both exclude the weighted mode outright) nor Theorem 5's weighted p-value. The unweighted rule fires hard (recall 0.52 on ISU-ILCC; "1st-true" = true first stops per bad cell, 0.044 at the frozen threshold 700) and every *mean* per-seed FSR sits inside $\alpha = 0.1$, but the per-seed ratio disperses (worst seeds 0.23–0.43, up to $33/100$ violating); one stream draw couples all of a seed's decisions. The "+ weighted decision" rows are the *paired* control: the same acquisition (same RNG and ordinal loaded-cell sequence, trace-audited; stopping shifts later load times, not the order) with the deployed weighted single-ratio decision (worst seed 0.059, zero violating, recall 0.02–0.03; the body paragraph reads the comparison). TONGJI's thin pooled supply closes the gate in all but $4/100$ seeds (recall 0.01, every first stop a true kill).

Corpus	Config	Recall	FSR	Worst	Viol.	1st (t/f)	1st-true
ISU-ILCC	m=24, 2 reps	0.39	0.059	0.43	21/100	47/11	0.013
ISU-ILCC	m=24, 3 reps	0.52	0.070	0.33	33/100	78/12	0.042
+ weighted decision	paired, same loads	0.02	0.003	0.06	0/100	13/4	0.007
ISU-ILCC	m=24, 3 reps, frozen 700	0.51	0.075	0.38	31/100	78/13	0.044
+ weighted decision	paired, same loads	0.03	0.002	0.06	0/100	16/3	0.009
MATR	m=16, 2 reps	0.18	0.014	0.23	5/100	20/6	0.005
TONGJI	m=12, 4 reps	0.01	0.000	0.00	0/100	4/0	0.002
TONGJI	m=12, 4 reps, frozen 350	0.01	0.000	0.05	0/100	4/0	0.002

1, conditional on the snapshot and $\{\tilde{L}_n \geq L_{\mathrm{spec}}\}$, the campaign's first realized stop kills n with probability at most $\left(\sum_j \alpha_j\right) \Pr[G_n^\circ]/\Pr[G_n] \leq \alpha/\Pr[\mathcal{C}_n \mid G_n^\circ]$, Theorem 2's factor again. Here, unlike there, it is 1 *on a set that can actually fire. A void index* burns *its slot for $\bar{c}$ instead of diverting it to a campaign load, so phase 0's channel releases, load-event count and campaign-RNG consumption through T_0 are identical in the two systems whether or not the stream voided; the first K phase-1 slots therefore load the* same *cells at the deterministic T_0 in both, before any decision exists and already holding the frozen calibration set, with all m stream indices resolved by T_0. For those the bound is exactly α. (Their admitted* records *do differ on a void, since H° ghosts where D burns, which is why a void triggers campaign-wide abstention, leaving no stop event there to transfer.) All K decide off one shared snapshot, so at most one of them is the campaign's first stop and thus covered. The guarantee is marginal per cell and marginal over the void event; nothing is claimed for decisions after the first stop, and a voided stream may not be redrawn, conditioning on $\{$no void$\}$ exits the i.i.d. calculus (at full exhaustion, $m = \sum_a s_a$, it is exactly the stratified counterexample; Example).*

Proof outline. Phase 0's assignment is resolved pre-campaign from the dedicated stream RNG and every slot is reserved for exactly $\bar{c}$ (deaths never free a channel, voids burn) so phase 1's start time and channel state are deterministic and no loading randomness depends on pool values. Up to and including the first realized stop the deployed system coincides pathwise with the no-stop reference (Theorem 2's coupling, simpler here: the void set is frozen at T_0), so every π_n realized in this regime is campaign-measurable and pool-independent. Conditional on the pool's goodness pattern the admitted values are i.i.d. with arm law π_{cal} restricted to good draws (Lemma 3's coordinate conditioning; Lemma 2 absorbs the reservation). I.i.d. arms make the joint density symmetric-times-single-ratio, so Lemma 4 applies with the deployed weights (the goodness tilt is a common constant the scale-invariant p-value absorbs) and the *with-replacement* arm draw is exactly what escapes the Example: without replacement the admitted composition is hypergeometric-dependent and no such factorization exists (the k-per-arm family violates at $\Pr[p \leq \frac{k+1}{2k+1}] = \frac{1}{2} + \frac{1}{2(k+1)}$ for every k). Void indices are handled by Theorem 1's ghost-value augmentation, arms are never redrawn. The first-stop containment, the common L_{spec} anchor on both sides (Lemma 5(i), an exact match here, not a relaxation) and gate monotonicity (Lemma 5(ii)), and the union bound over the cell's looks finish, with Theorem 2's factor on the transfer. Phase separation and void-*burning* are what make it 1 on a non-empty firing set. A void reserves its slot for $\bar{c}$ exactly as a supplied index does, so channel

releases, the number of load events, and the campaign RNG's consumption through T_0 carry no supply information: D and H° load the same K cells at T_0, which is $\mathcal{C}_n$ for those slots however the stream voided. They precede every decision while their eligible record set is already the full m (admission times are $\leq T_0$, load times $< T_0$). On a void the certifier abstains campaign-wide, so the stop event is empty there and nothing needs transferring. Sizing is pre-campaign arithmetic. The void probability is a multinomial tail in the supply counts, traded against the weighted granularity floor by the choice of m; the stream's i.i.d.-ness, which the exactness requires, is itself priced. □

Lower bounds: the prices are intrinsic. The three propositions below scope each price to a procedure class, not to our constructions. Two instantiate classical facts: Proposition 5 is the two-point total-variation bound of testing theory (see, e.g., Lehmann & Romano, 2005) applied arm by arm under the firewall, and Proposition 7 is the rank granularity of conformal p-values (Vovk et al., 2005). Proposition 6 quantifies, at every replication depth, the observation of Wieczorek (2023) that covariate-shift results do not carry over to stratified sampling.

Proposition 5 (Power is priced by prefix separation)**.** *Under Assumption 1, let R be any stopping rule whose stop event for cell n is measurable with respect to $(\mathrm{prefix}_{c_J}(V_n), a_n, W_n)$ with $W_n \perp V_n \mid a_n$ (the firewall structure; every mode in this paper), and let R be distribution-free valid at level α on a class $\mathcal{P}$. Fix per-arm good laws Q_g^a and bad laws Q_b^a whose half-half mixtures lie in $\mathcal{P}$ on every arm. Then*

$$\Pr[R \text{ stops } n \mid n \text{ bad}] \;\leq\; \alpha + \max_a d_{\mathrm{TV}}\big(\mathcal{L}(\mathrm{prefix}_{c_J}; Q_b^a),\, \mathcal{L}(\mathrm{prefix}_{c_J}; Q_g^a)\big).$$

In particular, when bad and good prefixes are indistinguishable at the looks, every *valid rule's certified recall is at most α; power beyond that is bought only by violating validity.*

Proof outline. With half-half mixtures on every arm, goodness is independent of (a_n, W_n), so the two conditional laws integrate over the same weighting; conditional on (a_n, W_n) the stop event is a fixed prefix set, and V_n's conditional law is the arm's mixture component, so the good/bad conditional stop probabilities differ by at most the arm's prefix TV distance; validity caps the good side at α. (The all-arm mixture is essential. Marginal validity does not cap arm-conditional good-stop rates, a rule concentrating its spend on one arm reaches *arm-conditional* bad recall $\min\{1,\ \alpha/\Pr[a_n = a \mid \mathrm{good}]\}$ with identical prefixes.) □

Proposition 6 (Composition rigidity)**.** (i) *For every $k \geq 1$, two arms with k reserved replicates each, equal single-ratio weights, and a frozen misranking forecaster admit within-class laws under which the deterministic full-exhaustion composition yields $\Pr[p \leq \frac{k+1}{2k+1}] = \frac{1}{2} + \frac{1}{2(k+1)} > \frac{k+1}{2k+1}$: no distribution-free* superuniformity certificate *exists for such compositions at any replication depth ($k = 1$ is the Example; a partial draw (two slots from a four-cell pool) violates at $25/36 > 2/3$; an M-arm chain violates at thresholds $\frac{k+1}{Mk+1} \approx 1/M$ with excess $\frac{k+2-M}{M(k+1)(Mk+1)}$ whenever $M < k+2$, e.g. $0.1091 > 0.1089$ at $M = k = 10$).* (ii) *Relative to the single-ratio calculus (pre-registered per-slot weights $\pi_n(a)/\pi_{\mathrm{cal}}(a)$ with π_{cal} a known pmf, exact superuniformity required) every composition family we exhibit fails: deterministic in supply, without-replacement exchangeable, nondegenerate two-point mixtures of i.i.d. streams ($\Pr[p \leq \frac{1}{k+1}] = \frac{1}{4} + \frac{1}{2(k+1)}$, exceeding it iff $k \geq 2$), and streams conditioned on $\{$no void$\}$ at full exhaustion ($m = \sum_a s_a$). I.i.d. draws from the known pmf support exactness (Theorem 5) and then face the trade: m at most the thinnest π-supported supply (at deployed replicate counts, granularity-starving, and widening support to thin arms only relocates the price to out-of-support abstention) or a strictly positive certificate-void probability (per cell or per campaign, per the design). Estimating π_{cal} exits exactness (Proposition 1's ρ^2); leaving the single-ratio calculus is the arm-matched escape at its own floor (Proposition 3); non-exchangeable pre-registered composition laws outside these families are not claimed.*

Derivations for Proposition 6. (i) Under the misranking forecaster all arm-1 good scores fall below all arm-2 scores, and within an arm the k records and the test's (boundary-substituted) score are i.i.d. continuous, so the test's within-arm rank R is uniform on $\{0, \ldots, k\}$. With all weights equal, a test on arm 1 gets $p = (1 + R)/(2k + 1) \leq (k + 1)/(2k + 1)$ always, while a test on arm 2 gets $p = (1 + k + R)/(2k + 1)$, which meets the threshold only at $R = 0$, with probability $1/(k + 1)$. Averaging over the equiprobable arms gives $\frac{1}{2} + \frac{1}{2(k+1)}$, exceeding $(k + 1)/(2k + 1)$ by $k/[2(k + 1)(2k + 1)] > 0$. The M-arm chain replaces the two blocks by M ranked ones: the

threshold $(k+1)/(Mk+1)$ is met by all of arm 1 and by arm 2 only at $R=0$ (arms $j \geq 3$ never), giving excess $(k+2-M)/[M(k+1)(Mk+1)]$, positive iff $M < k+2$. The partial draw is the same computation over the hypergeometric composition of two slots from four cells: $\frac{17}{18} \cdot \frac{1}{2} + \frac{4}{9} \cdot \frac{1}{2} = \frac{25}{36} > \frac{2}{3}$. *(ii)* Each listed family is refuted by (i) or by its own instance: deterministic-in-supply and without-replacement compositions are (i) verbatim; at full exhaustion, conditioning on $\{\text{no void}\}$ forces $N_a = s_a$ for every arm, the deterministic composition of (i) (away from exhaustion the conditioned law is a truncated multinomial, no longer i.i.d., and no certificate is claimed either way; when $m \leq \min_a s_a$ the conditioning is vacuous and Theorem 5's exactness stands); the two-point mixture of all-arm-1 and all-arm-2 streams gives $\Pr[p \leq \frac{1}{k+1}] = \frac{1}{4} + \frac{1}{2(k+1)}$, which exceeds $\frac{1}{k+1}$ for $k \geq 2$. That i.i.d. draws *do* support exactness is Theorem 5; the trade then follows from the multinomial law of the draw counts (N_a) against the supplies (s_a): $\Pr[\exists a : N_a > s_a] > 0$ whenever $m > \min_a s_a$ over π-supported arms. □

Proposition 7 (Granularity floor for rank rules)**.** *Let T be a non-randomized stopping statistic, symmetric in n good-only calibration records that are* exchangeable with the test cell's own score under the good-conditional law*, and invariant to strictly increasing transformations of the score axis, with ties occurring with probability zero (continuous marginal laws alone do not exclude ties among exchangeable scores). Then its good-conditional stop probability equals $|\mathcal{R}|/(n+1)$ for a fixed rank set $\mathcal{R}$, in $\{0\} \cup [1/(n+1), 1]$: no rank rule spends sub-granular budget, and Proposition 2(i) is the conformal instance. (Exchangeability is what makes the test's rank uniform; it is supplied by the ε-stream, and its absence is precisely what Proposition 1's ρ^2 and the Example below price.) The restriction to rank rules is necessary. A deterministic* composition *test (stop iff the test score falls in one region and all n records in another) is, for i.i.d. scores, distribution-free valid at level $\frac{1}{n+1}\left(\frac{n}{n+1}\right)^n < \frac{1}{n+1}$ (i.e. $\approx e^{-1}/(n+1)$), but its realized spending is nature's choice, never distribution-free power, and it remains capped by Proposition 5.*

Proof. Invariance to strictly increasing transforms reduces T to a function of the rank pattern of the $n+1$ scores; symmetry in the records reduces it further to the test score's rank $R \in \{0, \ldots, n\}$, so the stop region is a fixed rank set $\mathcal{R}$. Exchangeability of the test score with the records under the good-conditional law makes R uniform on $\{0, \ldots, n\}$ (ties have probability zero by assumption), so the good-conditional stop probability is $|\mathcal{R}|/(n+1)$. □

Example (the deployed weights admit no certificate here). Two arms, one reserved replicate each ($\pi_{\text{cal}} = (\frac{1}{2}, \frac{1}{2})$), $\pi_n = (\frac{1}{2}, \frac{1}{2})$: every weight in (2) equals one and $p = (1 + \#\{s_i \leq \hat{s}\})/3$. Let the frozen forecaster misrank the arms (arm-1 good scores in $(-11, -9)$, arm-2 good scores in $(9, 11)$, boundary scores tight) then a test on arm 1 yields $p \in \{1/3, 2/3\}$ and on arm 2 $p \in \{2/3, 1\}$, each with probability $\frac{1}{2}$, so $\Pr[p \leq 2/3] = 3/4 > 2/3$: the single-ratio weighted p-value is *not* superuniform under the stratified design. Randomization does not repair it. With atomless scores the tie-smoothed p-value $(U + \#\{s_i \leq \hat{s}\})/(2k+1)$, $U \sim \text{Unif}(0,1)$, violates at the same threshold with the same probability, the events $U + N \leq k+1$ and $N \leq k$ coincide almost surely, so the excess $\frac{1}{2} + \frac{1}{2(k+1)} - \frac{k+1}{2k+1}$ is composition-driven, not tie-driven. Validity there belongs to the arm-matched statistic alone, and obtaining i.i.d. calibration arms, as the ε-stream does by construction, is exactly what restores the single-ratio form of Lemma 4.

Open conjecture (unrestricted reuse). For decisions after the first stop under a shared calibration pool, conditioning on past stops tilts the calibration law and no exact finite-sample statement is available; we conjecture an $\text{FSR} \leq \alpha + \Delta$ bound with Δ controlled by the information content of the stop record, in the spirit of Barber et al. (2023), and rely on the runtime audit (Section 5) as the binding evidence at this generality.

B Extended related work

Battery prediction platforms. BatteryML (Zhang et al., 2024) and early-feature lifetime models (Li et al., 2024) standardize and improve cycle-life *prediction*. They sit one decision upstream of when a running test may be killed, and at what certified rate good cells are destroyed.

Early stopping in machine learning. Successive halving, Hyperband, ASHA, and BOHB (Jamieson & Talwalkar, 2016; Li et al., 2018; 2020; Falkner et al., 2018) assume budgets divisible

into informative increments with survivors extendable, as do DEHB (Awad et al., 2021), population-based training (Jaderberg et al., 2017) and the deep-kernel racer DyHPO (Wistuba et al., 2022), which extend or continue survivors; freeze-thaw (Swersky et al., 2014; Rakotoarison et al., 2024) resumes paused runs; learning-curve extrapolation (Domhan et al., 2015) stops irreversibly on a terminal loss predicted from a partial curve (our censored hitting-time labels are instead in the lineage of conformalized survival analysis, Candès et al., 2023). All fail for cells, which cannot be rerun once stopped and whose partial trajectories barely rank. We do not evaluate on learning-curve benchmark suites because their runs are resumable and labelled by terminal loss, so the irreversible false stop of a censored hitting time has no counterpart there. Racing (Maron & Moore, 1993; Birattari et al., 2002) needs repeated cheap i.i.d. evaluations (impossible with one irreversible trajectory each) while conformal solver stopping (Clarke & Stellato, 2026) calibrates one threshold over whole solve trajectories, so needs no multiple-look correction; an unreplayable campaign pays for its looks, our α-spending. That spending is a pre-frozen budget split across looks, valid by a union bound, in the tradition of Lan & DeMets (1983) (whose boundaries instead exploit the joint law of the test statistics).

Conformal prediction under adaptive data collection. Conformal prediction (Vovk et al., 2005); its weighted extension (Tibshirani et al., 2019) handles known density ratios, Fannjiang et al. (2022) extends it to feedback covariate shift (efficient jackknife+ variants: Prinster et al., 2023), and Prinster et al. (2024) certifies validity under *any* agent-induced shift with known query densities (tractable in structured cases, factorially many weight terms in general); Barber et al. (2023) bounds nonexchangeability's cost and Jin & Candès (2023) controls FDR with a calibration set exchangeable with the test units; hierarchical and class-conditional conformal methods (Lee et al., 2026; Ding et al., 2023) calibrate with few samples per group, the regime in which our arm-matched (Mondrian) p-value of Proposition 3 abstains, and extending them to adaptive acquisition is open; off-policy, BO, and online variants (Taufiq et al., 2022; Stanton et al., 2023; Gibbs & Candès, 2021) target prediction coverage or long-run rates, none give per-decision control for irreversible kills; ours obtains a clean calibration set by design. Our decision statistic is not new. The weighted, one-sided conformal p-value for "is this unit's response past a user-specified threshold" under covariate shift is exactly Jin & Candès (2026); the inlier-only calibration reference is the construction of Bates et al. (2023); and Nair & Janson (2026) give exactly-valid randomization tests on adaptively collected data. What is new here is not the statistic but the design that makes its weights exactly known *inside a live campaign*, and the price theory for what that costs.

Confidence-bound stopping. Xiao et al. (2026) stops single experiments by classifying them against a threshold with a GP confidence bound, certified by a concentration bound on cross-validated out-of-fold rates under i.i.d. evaluations. The regime here differs on every axis the theory uses, right-censored hitting times, adaptive acquisition, and one calibration set reused across many decisions, so we do not run it as a baseline; the closest rows in our replay are the tuned-margin threshold and, for a concentration-bound certificate, the fairly scheduled LTT.

C Extended experimental results

Scope and negative results. Pseudo-observation feedback of stopped trials into the surrogate (*soft rejection*) follows Stenger et al. (2024), ablates neutral in our finite-pool replays, and we claim nothing for it. An anytime e-process variant (Ramdas et al., 2023) with power calibrators $f_\gamma(p) = \gamma p^{\gamma-1}$ (Vovk & Wang, 2021) certifies nothing at realistic calibration sizes ($f_{0.5}$ needs $p \approx \alpha^2/4$ against a granularity floor of $1/(|\mathrm{Cal}_j| + 1)$): a negative result, reported in the *rule_b* rows below, and α-spending is the right tool at $|\mathrm{Cal}| \approx 30$. We claim no pool-level false-discovery control and run none; e-BH (Wang & Ramdas, 2022) is the natural candidate, since it tolerates arbitrary dependence.

Denominator convention under truncation. Rates in the budgeted regime are taken over cells *resolved* before the cut: a cell still on a channel when consumption reaches the budget is charged its partial occupancy and yields no outcome, so it enters neither numerator nor denominator. Everywhere else the denominator is the loaded-cell one of Section 5, and the two coincide whenever no truncation occurs, which is every unbudgeted campaign and so the whole of Table 2. Under truncation at most $K = 8$ cells sit on channels at the cut, so a budgeted seed reporting f false stops over n resolved good cells has loaded-cell rate at least $f/(n + 8)$. The single exceedance we

report falls inside that gap. The worst weighted MATR seed is $4/37 = 0.108$ resolved and at least $4/45 = 0.089$ loaded, so whether it exceeds α is a matter of convention; the worst ISU-ILCC seed moves from $3/52 = 0.058$ to at least $3/60 = 0.050$. Recovering the exact loaded denominators needs the in-flight cell identities, which the shipped result files do not carry.

Budgeted deltas in full (Section 5). Across $\alpha \in \{0.05, 0.1, 0.2\}$ the static rule's mean FSR tracks well below its budget with recall monotone in α (Fig. 1(b,c); details below), with abstention where granularity cannot support it. The weighted sweep in numbers (cluster runs; common 50-seed base, the α=0.1 point being the 100-seed run restricted to those seeds, maxima over unequal seed counts are not comparable): recall rises $0.004 \to 0.045 \to 0.115$ (MATR) and $0.090 \to 0.219 \to 0.384$ (ISU-ILCC) across $\alpha \in \{0.05, 0.1, 0.2\}$, with worst-seed FSR at most 0.65α (the full 100-seed α=0.1 worst seed is 0.095); HUST's static α=0.2 worst seed (0.31) continues above Fig. 1(b). Across budgets the comparison is less favourable to the weights: on MATR the static rule at $\alpha = 0.05$ reaches recall 0.32 with worst seed 0.057 (50 seeds), against the weighted rule's 0.05 at $\alpha = 0.1$, so the weights buy worst-seed stability at a fixed budget, not recall at matched safety. Under a 40% channel-cycle budget (100 seeds; the budget is $0.4\times$ the pool's summed horizons, binds on MATR and ISU-ILCC, where it cuts no-stop occupancy by 15% and 11%, and is nominal on HUST, where the cell-mode calibration reserve already leaves the campaign-loadable pool at the budget and the cut is 0.2%), the acquisition-matched no-stop arm isolates what stopping buys. EI acquisition without stopping cuts *observed-endpoint regret* against uniform loading on ISU-ILCC ($40 \to 18$), while on MATR the cut ($162 \to 123$) is within noise (regret in Table 12, paired CIs in Section 5, completions-only convention in Table 11), and calibrated stopping then adds safety at no regret cost. Seed-paired deltas versus EI no-stop are null on ISU-ILCC ($+1$ $[-7, 9]$ static; $+4$ $[-5, 16]$ weighted) and HUST (-0.1 static; exactly zero for the abstaining weighted rule), and favorable but unresolved on MATR (-34 $[-77, 1]$; -24 $[-55, 1]$). Heuristic stopping destroys the campaign's product. ASHA observes *no* top-decile endpoint in $76/100$ ISU-ILCC seeds, its observed regret exploding $18 \to 2143$ versus EI no-stop (Table 12), while tested-arm regret is blind to the distinction. Seed-paired discovery under the static and weighted modes matches EI no-stop and is *earlier* on ISU-ILCC (-921 static, -935 weighted), favorable but unresolved on MATR (-115 cycles $[-230, 1]$ static), budgeted and unbudgeted alike (Table 11).

Limitations in full (Section 7). Every campaign here is an offline replay. No physical channel was occupied, and scheduling and channel contention are replayed from recorded trajectories. Campaign cells run to their recorded horizon, so busy-cycle savings are stated under the convention most favorable to stopping (corpus detail above). All five battery corpora cycle under prescribed laboratory protocols, so the lives (and with them L_{spec}) are protocol-specific (scope note below). The cross-domain evidence is one real corpus, IVPT, which breaches look-coverage on 2 of 135 runs. The power we report leans on the imposed cycling conditions. Under depth-of-discharge control the weighted rule's recall falls to 0.02.

Ablations in full (Section 5). The SCHED-optimized spending schedule contributes up to +50pp recall over a uniform split, but that gap is not a second, independent mechanism. On MATR the uniform comparator is never feasible at $|\mathrm{Cal}_j| \in [16, 23]$, so it degenerates to the whole budget at cycle 50, before the $\Delta Q(V)$ feature enters at cycle 100. What SCHED buys there is look selection *into* the $\Delta Q(V)$ regime, which is why the uniform and *nodqv* ablations land in the same place (0.032 and 0.045 against the seed-matched 0.529; the 100-seed base reads 0.536). The $\Delta Q(V)$ feature is the power source on MATR (without it, static recall collapses $0.529 \to 0.045$ and weighted $0.045 \to 0.004$ (seed-matched baselines)) yet immaterial on ISU-ILCC ($0.38 \to 0.38, 0.22 \to 0.22$), whose signal is carried by the imposed cycling conditions (*nodqv* rows in Tables 8–10). Deliberately mixed-pool calibration inflates FSR up to $2.6\times$ on HUST, breaching the worst-seed standard (rows below). The good-only restriction is provable *and* empirically binding. Breaking Assumption 1, degrading the forecaster, doubling reserved replicates, and splitting the pool by composition all leave the weighted mode safe or abstaining, while the uncovered static mode absorbs the damage (stress-probe paragraphs below).

Corpus detail and provenance (Section 5). *All five corpora's counts are our own recount of the replayed subset, retrieved through* BATTERYLIFE (Tan et al., 2025); *the releases' own totals differ, and the citations name provenance, not these counts.* MATR (Severson et al., 2019; Attia

et al., 2020) (169 lithium-iron-phosphate (LFP) cells over 77 distinct charge protocols (68 one- or two-stage $(\mathrm{CC}_1, \mathrm{SOC}_{\mathrm{switch}}, \mathrm{CC}_2)$ over 124 cells, 9 four-stage SOC-windowed over 45), 34 replicated; ISU-ILCC (Thelen et al., 2023)) 240 NMC pouch cells over 61 cycling conditions (charge rate $\times$ discharge rate $\times$ depth of discharge), 57 with four replicates and four with three; HUST (Ma et al., 2022), 77 LFP cells, one multi-stage discharge protocol each (single-replicate regime). Two generality corpora extend the span (Table 6; same recount convention): SNL (Preger et al., 2020) (52 cells in 19 chemistry-crossed arms) and TONGJI (Zhu et al., 2022), 99 cells, 10 arms, 2–26 replicates each. Cells with fewer than 110 recorded cycles are excluded (nine TONGJI cells of one arm with lives 18–27 cycles, which moves the TONGJI pool median from 325.5 to 369; no other corpus loses a cell), as are cells without a shipped label. Completion is the recorded test horizon, which on four corpora runs past end of life. Cycles after $\tilde{L}$ are 77% of pool channel-cycles on ISU-ILCC, 42% on TONGJI, 29% on SNL, 4.5% on MATR and none on HUST; busy-cycle savings and discovery deltas are therefore stated under the convention most favorable to stopping and can only shrink under completion at $\tilde{L}$. FSR is unaffected for the per-cell rules, and recall can move only through the at most five bad cells per corpus that are past end of life at a look while still running. Arms are urns of real recorded cells drawn without replacement; $K = 8$ channels, $\varepsilon = 0.2$, looks $\{50, 100, 200\}$, fixed L_{spec} at the pool-median life in recorded cycles (on ISU-ILCC these are BATTERYLIFE's 80%-SOH labels (the release ships none), spanning DOD levels; equivalent-full-cycle (EFC) and frozen-threshold replays: below (safety survives both; cross-DOD power does not survive EFC relabeling)) $\alpha = 0.1$; 100 seeds for the main table, deployment audits, ε-sweep, dose, frozen-threshold, held-out-composition, budgeted, equivalent-full-cycle and specification-stress arms, 50 for the α-sweep, ablations, appendix baselines, η reruns, the degraded-forecaster, no-$\Delta Q(V)$ and batch-holdout probes and the L_{spec} and depth-of-discharge sensitivity replays, 20 for the solid-state cohort; seed-block bootstrap confidence intervals (CIs). Data provenance (per-file SHA-256, repo revisions) and the preprocessing audit (including the $\Delta Q(V)$ check against Severson et al. (2019) (corr. -0.87 on our 169-cell pool)) ship with the code.

When does the weighted rule have power? The recall spread (0.22 ISU-ILCC, 0.14 TONGJI, 0.05 MATR; abstention on HUST; two stops in 100 seeds on the 52-cell SNL) is predicted by three pool statistics available *before* a campaign: replicate count per arm (three to four on ISU-ILCC; one on HUST), the overlap between pool support and the acquisition's focus ($61/61$ arms on ISU-ILCC, $34/77$ on MATR), and the weights' effective sample size; baseline tuning follows below.

What the budget buys. Of an ISU-ILCC campaign's 54 eventual winners, halving destroys 52, naive thresholding 10 (worst seed 30 of 60), the tuned margin 1.6, the weighted rule 0.5 (at most 3); on HUST naive's worst seed destroys the entire 19-winner pool. A screening campaign exists to find exactly these cells.

Tables 8–10 list every (experiment, policy) pair of the canonical sunk-convention campaign with seed-block bootstrap 95% CIs and per-experiment violating-seed counts, generated directly from the released result files with no hand transcription (the extension suites (the weighted α-sweep behind Fig. 1(b,c), the six-corpus LTT runs, and the budgeted racing/LTT arms) are cluster runs of the same image and environment, shipped in the artifact and summarized in the text rather than tabulated here); deployment-audit rows appear under the *theorem* experiment names, and the *nodqv* rows ablate the $\Delta Q(V)$ feature (forecaster retrained without it) on the two datasets that use it. Table 11 reports the operational costs (discovery time, busy channel-cycles). Budget-limited and case-study numbers are quoted in Section 5 from the released pilot result files, which ship with the code together with the per-dataset horizon audit. All intervals are Monte Carlo over partition and scheduling seeds at fixed corpus; they do not model sampling from a wider cell population.

The worst-seed excursion of the unweighted mode in Fig. 1(b) reaches 1.54α on HUST at both $\alpha = 0.1$ and $\alpha = 0.2$ (50 seeds); the 100-seed run of Table 2 puts the same mode at 2.67α. This is the unweighted mode's finite-calibration dispersion. Prefix recalls and worst-seed ratios for the weighted reruns are below.

Censoring audit (Section 2). Determinacy exceptions, cells censored alive before the spec with original labels beyond it: zero on MATR, ISU-ILCC, SNL and TONGJI; on HUST, two of 77 cells (`HUST_1-3`, `HUST_10-3`) end their records 15–25 cycles short of $L_{\mathrm{spec}} = 1873$ with labels 1911 and 1901; the convention counts them as not-observed-good. Every HUST record ends at or before its shipped label, at a last recorded capacity of 0.891–0.906 Ah (81–82% of the 1.1 Ah nominal), so

$\tilde{L} = C_x$ for all 77 cells and $L_{\text{spec}} = 1873$ is the median record length; ranking is preserved because the records end at a near-constant state of health, and only the two cells above can flip. Stated bias: recall denominators gain, and FSR denominators lose, at most those two cells; on any seed where one of the two is stopped, per-seed FSR is correspondingly (negligibly) under-counted.

Burn-in stream sizing. Table 4 sweeps the front-loaded stream's size m (Theorem 5). The design rule is pre-campaign arithmetic. The void probability is a closed multinomial tail in the supply counts, and it competes with the weighted granularity floor. On ISU-ILCC (two reserved replicates across 61 arms) pilot recall peaks at moderate m, 0.125 at m=28 against 0.113 at the deployed m=24, and collapses to total abstention by $m = 56$; TONGJI's uneven replication needs deep reserved urns (4 per arm) before the stream fires in more than a handful of seeds (3–4 of 100 at two, 9–29 at four). On the certified first stop, three replicates read 0.026 (a halved bad-cell universe, not more kills) and TONGJI reads 0.012; an uncovered composition-proportional audit arm reaches 0.044 (frozen spec), mean-safe but worst-seed-fragile, unlike the weighted audit (Tables 4/3). The deployed burn-in's post-first-stop worst seeds reach 0.14, its recall is 0.11 against the replicate-matched weighted rule's 0.19, and the thin-supply corpora certify little (MATR at 0.003–0.009; HUST abstains at every m).

Table 4: Burn-in stream sizing (Theorem 5): the i.i.d.-ness the exactness requires is itself priced. Larger m buys calibration mass against the granularity floor but raises the *void* tail $\Pr[\exists a : N_a > s_a]$ (computed exactly from the multinomial law; any void aborts the certificate for the whole campaign), so the two columns move against each other and recall peaks at moderate m. On ISU-ILCC (supply 2 across 61 arms) $m = 24$ already voids 36% of campaigns, and by $m = 56$ the certificate is essentially never issued. "fires" counts seeds with any stop, so it mixes voided campaigns with granularity abstention. Recall is averaged over *all* seeds, voided ones contributing zero (conditional on firing, ISU-ILCC m=24 recall is 0.21). FSR is the per-seed mean (max seed in parentheses), matching Table 2's column rather than the pooled primary metric of Section 5; the two differ by at most 0.001 on these rows. FSR and the first-kill rate are full-campaign audits; only the per-cell first-stop risk is theorem-controlled (0.0023 at m=24), the excess beyond it being calibration-reuse inflation (the Remark after Prop. 3). "1st (t/f)" counts campaign-first stops (the only stops the certificate covers) as true/false kills: at ISU-ILCC m=24 they are $47/8$ of 470 total stops, an audited all-stop recall of 0.11 against a covered first-stop recall of 0.013. The † row runs the same TONGJI design at the frozen threshold 350 (pool median rounded down to 50, fixed before the seeds ran; a fixed-threshold replay, not an external specification): 29 campaigns fire, every first stop is a true kill, and the frozen-threshold first-stop audit recall is 0.014 at mean FSR 0.004. The thin-supply corpora show the same arithmetic from the other side. MATR reserves a replicate on only 34 of 77 arms and HUST is single-replicate throughout, so their supports admit no sizing that both clears the granularity floor and keeps the void tail small; MATR stays safe at negligible recall and HUST abstains outright at every m tried. All rows are 100 seeds; the per-arm reserved-supply multisets behind each $\Pr[\text{void}]$ are partition-determined and ship with the artifact.

Corpus	m	reps	seeds	$\Pr[\text{void}]$	fires	1st (t/f)	recall	FSR (max)	first-kill
ISU-ILCC	24	2	100	0.36	55/100	47/8	0.113	0.020 (0.143)	0.15
ISU-ILCC	24	3	100	0.05	56/100	48/8	0.067	0.012 (0.136)	0.14
ISU-ILCC	28	2	100	0.51	44/100	40/4	0.125	0.025 (0.188)	0.09
ISU-ILCC	32	2	100	0.65	29/100	27/2	0.092	0.017 (0.182)	0.07
ISU-ILCC	40	2	100	0.87	12/100	11/1	0.047	0.006 (0.158)	0.08
ISU-ILCC	56	2	100	1.00	0/100	—	0.000	0.000 (0.000)	—
TONGJI	6	2	100	0.20	3/100	3/0	0.002	0.000 (0.000)	0.00
TONGJI	10	2	100	0.67	4/100	4/0	0.003	0.000 (0.000)	0.00
TONGJI	10	4	100	0.16	18/100	16/2	0.013	0.001 (0.045)	0.11
TONGJI	12	4	100	0.27	26/100	26/0	0.028	0.002 (0.105)	0.00
TONGJI	12†	4	100	0.27	29/100	29/0	0.030	0.004 (0.105)	0.00
TONGJI	14	4	100	0.41	28/100	26/2	0.026	0.002 (0.050)	0.07
TONGJI	16	4	100	0.57	25/100	22/3	0.032	0.005 (0.105)	0.12
TONGJI	20	4	100	0.86	9/100	9/0	0.011	0.001 (0.056)	0.00
MATR	8	2	100	0.21	11/100	10/1	0.003	0.001 (0.031)	0.09
MATR	16	2	100	0.72	24/100	13/11	0.009	0.005 (0.083)	0.46
HUST	8	1	100	0.68	0/100	—	0.000	0.000 (0.000)	—
HUST	16	1	100	1.00	0/100	—	0.000	0.000 (0.000)	—

Campaign-level incidence. The guarantees are per-trial and marginal (Section 4); a small per-cell false-stop probability across many good cells still accumulates campaign-level incidence, which we

report rather than control. Weighted rule, 100 seeds: MATR stops in 73 campaigns, with a false first kill in $5/73$ and at least one false stop in $14/100$; ISU-ILCC reads 100, $4/100$ and $35/100$; HUST abstains, $0/100$. The online mode across the 600 theorem-mode campaigns: at least one false stop in $22/600$. Campaign-wise control is an explicit non-goal (first paragraph of this appendix); e-BH over per-cell e-values is the natural extension.

Held-out seed bank. As a final audit, every corpus's main and weighted arms were replayed at 50 *fresh* seeds (offset 1000) never used during development. Every qualitative conclusion reproduces: heuristic violation magnitudes (ASHA 0.38–0.95, naive 0.12–0.60), the static mode's mean-safety with worst-seed fragility (max 0.20 on HUST), the weighted rule's worst-seed safety on all five battery corpora (0 violating seeds; recall 0.22 ISU-ILCC, 0.12 TONGJI, 0.03 MATR) (and the cross-corpus breaches again land on IVPT, under both look conventions: noncausal $1/50$ (max 0.111) and causal $\{4, 8, 22\}$ h weighted $1/50$ (max 0.148; recall 0.19, static $2/50$, max 0.194)) consistent with the audit-only status of the weighted mode on this corpus. The bank defends against seed-level selection only; configuration-level choices (which constants became headline rows) are a disclosure matter (the power-boosted rationale below, the burn-in m-sweep shown in full above) since all design iteration shared these five corpora.

The running pair, replayed. At cycle 100 the introduction's two cells are indistinguishable by capacity rank, so halving kills whichever lands in the losing bracket; the *static* rule holds both until the $\Delta Q(V)$-driven margin separates them, the *weighted* rule chiefly because its gates rarely fire.

Two further probes. Arm-level racing (retire arms whose completed replicates all miss the spec) controls false elimination trivially but releases no channel time (Table 11); sweeping $\varepsilon \in \{0.05, \ldots, 0.5\}$ leaves online shared-pool recall ≈ 0.01 or below everywhere. With one reserve per design the with-replacement stream meets an emptied arm after about nine successful draws on ISU-ILCC and seven on MATR (in expectation), roughly half of them good, against a five-record cold start, after which the shared mode abstains for good; reserve depth and certification latency ($\bar{c} = L_{\mathrm{spec}}$) bind, not the stream share. Two reserves per design raise recall to 0.03 (the power-boosted arm).

Depth-of-discharge sensitivity (ISU-ILCC). On the cycle axis, ISU-ILCC goodness is strongly condition-driven. Over the $46/61$ conditions whose processed metadata resolves a total discharge SOC swing (the remaining 15 carry degenerate stage bookkeeping and are excluded here), Spearman's $\rho(\mathrm{life}, \mathrm{DOD}) = -0.85$ and $\Pr[\mathrm{good} \mid \mathrm{DOD} \leq 0.45] = 0.95$ versus 0.04 at $\mathrm{DOD} > 0.7$, these two against the median life of the 181 resolvable cells (925 cycles), the probe's own spec; against the campaign's $L_{\mathrm{spec}} = 721$ the same contrast reads 0.96 versus 0.20. An equivalent-full-cycle relabeling overlaps the cycle-axis good set at Jaccard 0.70 (0.68 at L_{spec}). This is the disclosed signal structure (the imposed conditions, DOD included, are the forecaster's design vector) and the safety claims do not lean on it. Replaying *within* fixed-DOD strata ($[0.40, 0.65]$, 69 cells; $[0.70, 0.95]$, 56 cells; 50 seeds, within-stratum median spec) reproduces the ladder's ordering exactly, ASHA at FSR 0.75 and naive at 0.46–0.49 with near-total worst-seed violation in both strata, against calibrated means of at most 0.013; the weighted rule has zero violating seeds in both, the static rule one in the mid stratum (max 0.118), and both retain only residual within-stratum power: static recall 0.09 mid and 0.004 deep, and for the deployed weighted rule 0.021 and 0.0026. Under condition control the flagship's power is essentially gone; what survives is its safety. Safety is DOD-invariant; certified cross-DOD power rides on the condition signal, as stated in Section 5. The full equivalent-full-cycle replay quantifies it: relabeling life as $\tilde{L}\times$ the group's total discharge swing (the shared `measured_total_dod`; the 15 degenerate-bookkeeping conditions are excluded, leaving 181 cells on 46 arms; cycle-axis looks unchanged; 100 seeds), safety holds (the weighted rule keeps zero violating seeds (max 0.077) while ASHA (0.84 mean FSR) and naive (0.49) violate on every seed) but power does *not* survive the relabeling: weighted recall falls to 0.05 (static 0.07, max 0.122) from 0.22 (0.39) on the cycle axis. The cross-DOD condition signal carries most of this corpus's stopping power; a deployment whose objective is charge-throughput life must budget for that loss, and the safety guarantees are indifferent to the choice. All five battery corpora replay prescribed laboratory protocols (constant or multi-step current profiles, never a dynamic duty cycle), so the recorded lives, and with them L_{spec}, are protocol-specific. Over 92 cells, Geslin et al. (2025) measure realistic loads extending life by up to 38% against constant current. We claim no transfer of

absolute lifetimes; the guarantee is conditional on whatever life definition a campaign adopts, and the relabeling above measures what changing that definition costs, safety unchanged, power reduced.

Specification-threshold sensitivity. The replay convention sets L_{spec} to the pool median (Section 2). Re-running MATR at $L_{\mathrm{spec}} = 608$ and 912 ($\pm 20\%$; 50 seeds) keeps the safety ordering stable while stopping power is specification-sensitive. The weighted rule stays worst-seed-safe at both thresholds (max-seed FSR 0.064 and 0.080, $0/50$ violating seeds) at sharply reduced recall (0.018 and 0.011, versus 0.045 at the median), and the unweighted static rule remains mean-safe (FSR 0.009 and 0.015; recall 0.265 and 0.279, versus 0.529; both median baselines on the same 50 seeds) but shows its characteristic worst-seed fragility at the harder spec ($3/50$ seeds above α, max 0.190 at 912). Safety ordering is stable; certified power is not, consistent with the support and granularity preconditions of Sections 3.2 and 5. A frozen threshold removes the seed-level pool statistic: at round constants frozen before the runs, 750/700/1800 for MATR/ISU-ILCC/HUST against pool medians 760/721/1873, each the median rounded down to a round constant (50 on MATR and ISU-ILCC, 100 on the longer-lived HUST scale); fixed-threshold replays, not external specifications; 100 seeds. The ladder replicates: weighted 0.051/0.21/abstains (max seeds 0.093/0.065/none, zero violating), online shared 0.004/0.029/abstains, and the front-loaded burn-in on ISU-ILCC (m=24) audited recall 0.12 with $47/9$ covered first-stop kills, while the uncovered static mode breaches on all three (worst seeds 0.140/0.125/0.125, 1/1/2 of 100) exactly as its audit-only status predicts. The covered online arm is not immune at the seed level either. Its ISU-ILCC worst seed is 0.114 ($1/100$), inside the power-boosted family's disclosed 0.158 maximum and outside the per-cell first-stop quantity the theorem controls. Every quantity in the theorem arms' statements is thereby fixed pre-campaign, a frozen-threshold replay audit under the exact algorithmic configuration; freezing removes seed-level threshold adaptivity but, since the rule reads the same pool's endpoints, does not make the spec exogenous.

Confirmatory specification stress test (Lemma 6). Table 5. The one protocol in this paper whose specification is campaign-independent: a disjoint specification split (non-declared-arm cells, excluded from every campaign partition) fixes L_{spec} by a rule frozen before the one-shot run, and the equal-urn pre-registered-sequence construction (the mode Lemma 6 covers at the tested-cell estimand) is run once against a matched never-stopping control. Result: total abstention on ISU-ILCC at $L_{\mathrm{spec}} = 850$ (zero stops in 100 seeds, zero resource delta), and on SNL the split's $L_{\mathrm{spec}} = 300$ leaves no eligible bad test cell in the declared support (criterion ineligible). Pool-median power did not transfer to this independent specification protocol; certified power under this construction is highly sensitive to specification choice. This is a falsification-style stress result, not a deployment demonstration; exploratory pool-median pilots of the same construction (which do fire, at recall up to 0.57, with per-seed FSR ratios dispersing to 0.50 as Lemma 6 permits) remain development-set observations and are not reported as covered operating points.

Table 5: Confirmatory specification stress test (100 seeds, one-shot frozen protocol): the equal-urn pre-registered-sequence construction of Lemma 6, with L_{spec} set by a *campaign-independent disjoint specification split* (non-declared-arm cells, excluded from every campaign partition; rule and constants frozen before the run). On ISU-ILCC the construction abstained entirely, and the matched never-stopping control shows the resource delta is exactly zero. The SNL row is a *pre-run feasibility screen*, not a confirmatory run. Its frozen spec leaves no eligible bad test cell in the declared support (criterion ineligible), so the leg was never run and its stop count is unmeasured. What this costs in transfer is discussed with the row (App. C).

Corpus	SPEC source	L_{spec}	Eligible bad	Stops	Recall	Busy saving
ISU-ILCC	disjoint non-declared-arm split	850	2140 (pooled)	0	0.00	0
SNL	same protocol	300	0	— (not run)	N/A	N/A

On Theorem 2's coupling factor. The shared-pool bound carries $\Pr[G_n^{\circ}]/\Pr[G_n]$, which is 1 when $\mathcal{C}_n$ (deployed and reference load the same cell at slot n, which only a preceding stop or empty-supply draw can spoil) is sure. We report it as a stated price and do not measure it, for two reasons worth recording.

First, the two spoilers fire on different schedules. Stops are rare here. Over 100 seeds the shared-pool arm realizes no stop at all in 100 of 100 campaigns on HUST, 99 on MATR and 80 on ISU-

ILCC (`scripts/measure_coupling_factor.py` logs each campaign's first-stop load index). Supply exhaustion is not rare, and it also desynchronizes the two systems, because an exhausted draw loads a campaign cell where H° loads a ghost. The no-stop frequency above therefore does not bound the two spoilers jointly, and we do not report it as an estimate of the factor.

Second, estimating $\Pr[G_n^\circ]$ needs H° itself, whose defining feature is a calibration stream that never exhausts (value-only ghosting at the pre-registered arms). Replaying the same policy with `decide_stop` suppressed reproduces only the no-stop clause; it inherits the deployed run's exhaustion, so its profile is not H°'s. This replay is in any case already reported as not instantiating that ghosting, since the pool-median conditioning tilts the ghost values. Front-loading makes $\mathcal{C}_n$ sure on a set that fires, the K slots mounted at T_0, which precede every decision, and where burning a void slot rather than diverting it keeps the two systems on the same cells (Theorem 5). Beyond those K the stop spoiler is live again. A pre-registered mounting order, available only inside that same front-loaded design, pins slot n's cell at every slot, at the further cost of the adaptive acquisition; it still forfeits the certificate on a void, and under a binding channel-cycle budget the two systems can reach different numbers of slots even with the order pinned. For the shared pool the factor stays a stated, unmeasured price.

Replay convention for the offline pool. Every reported number uses the *historical* convention (the offline pool's cells are never loaded in the campaign and never charged (the ε-coin is still consumed, so stream alignment is preserved); their scores are read from the pre-campaign records). The convention is not cosmetic. The unweighted static rule's stop decisions are per-cell deterministic functions of the frozen snapshot and pool scores, so its safety and recall figures are invariant to it (verified per seed against the alternative convention, in which the pool is also cycled at share ε and charged). The weighted rule's are *not*: its weights read the acquisition pmf at the test cell's load time, which the convention shifts, as do all timing, busy-cycle, and order-sensitive heuristic figures. Only the theorem modes, whose $\bar{c}$-occupancy stream is purchased explicitly, are unaffected by construction.

Configuration disclosure. All deployed-rule constants are pre-campaign design choices, identical across policies within an experiment: $K = 8$ channels; looks at cycles $(50, 100, 200)$ with L_{spec} the pool median life; default calibration share $\varepsilon = 0.2$; acquisition softmax over EI with uniform floor $\lambda = 0.2$ (the theorem audits' power-boosted configuration uses $\varepsilon = 0.3$, $\lambda = 0.5$, two calibration replicates per arm, chosen during development, after replays at the default certified recall ≤ 0.01, to give the theorem modes a non-vacuous audit; it is a sensitivity configuration, not a pre-registered study arm, and its worst seed 0.158 is reported under exactly that label); the softmax temperature is *adaptive*, set at each selection to the standard deviation of EI across active arms plus 10^{-3} (scale-free, no tuning), and the realized pmf is recorded; π_{cal} proportional to initial calibration supply; block-exclusive audits use block size $m = 5$ with round-robin reserved supply. Partitions, in the following order: calibration first (at the configured reserve r, arm mode reserves $\min(r, n_a - 1)$ per arm and skips singletons; cell mode takes 35% of the pool, on HUST and the solid-state cohort), then, of the *remainder*, schedule-tuning 15% and forecaster 25%, the rest campaign urns. The per-corpus main configurations use $r \in \{1, 2\}$ and realize pool shares of calibration 19–36%, forecaster 15–20%, schedule-tuning 8–12% (the solid-state cohort supplies the calibration maximum and the schedule-tuning minimum); the burn-in, coverage-dose and composition arms reserve deeper (r up to 4), which on ISU-ILCC at $r = 3$ takes calibration to 75% and leaves schedule-tuning 4%, the resource price those rows report. The schedule optimizer searches all look-subsets, allocates uniformly within a subset subject to the granularity-feasibility constraint, and picks the subset maximizing expected released channel-cycles on the tuning partition, once, pre-campaign; theorem modes instead spend the whole budget on the final look. Generality corpora: SNL and TONGJI run without $\Delta Q(V)$ (protocol-confounded, the HUST convention); TONGJI reserves two calibration replicates per arm by the pre-campaign granularity arithmetic ($|\mathrm{Cal}_{\mathrm{good}}| \approx 6$ at one replicate, under $1/\alpha - 1$ on every seed); SNL keeps one because two are unaffordable at 52 cells (26 of them reserved), not because two would miss the floor. On SNL the weighted gates close almost entirely (two stops in 100 seeds) while the pooled unweighted mode clears its floor at recall 0.02 with worst-seed fragility (0.125; $2/100$). The coverage-dose arm (*mainw2* rows) doubles the reserved replicates per arm and buys no recall ($0.22 \to 0.19$ on ISU-ILCC). The weighted rule's precondition is support coverage, not replicate count. The held-out-composition row (*wsplit*) probes Proposition 1's gate. At one replicate per arm the composition and score halves have disjoint arm supports, so the weight gate closes

and the row abstains everywhere, the stratified pool does not realize the proposition's i.i.d. sampling model, which stands as theory, not as this row's certificate. The mixed-pool ablation's HUST breach in detail: $6/50$ violating seeds, max 0.235, versus $1/50$ for good-only. The theorem-mode conditional-on-stop record in detail: the first stop was false in 16.9% of stopping campaigns, 22.5% on power-boosted ISU-ILCC, where the 0.158 worst seed lives. The degraded-forecaster probe in detail: with $N(0, 0.15)$ log-life noise added to the forecaster's training targets only (evaluation and calibration labels untouched) the static rule's worst seed hits 0.15 $(1/50)$ and its MATR recall falls $0.53 \to 0.19$, while the weighted mode's worst seeds stay ≤ 0.061. The assumption stress in detail: leave-one-batch-out calibration on MATR sends the unweighted rule's worst seed to 0.30 ($3\times$ budget) while the weighted gates abstain on all three folds. Two limits of that evidence: holding out MATR's B2 leaves 0 good cells at the pool-median spec, so its zero false-stop rate is $0/0$ and a false stop is impossible there by construction; and B4 was not among the folds run.

The solid-state case study in detail. This 36-cell LFP|PEO@LLZTO|Li cohort (20 seeds, sunk convention, local run) runs on a shorter horizon than the public corpora, so it does *not* share their look grid or pool-median spec: looks are $\{30, 60\}$ cycles with a fixed $L_{\text{spec}} = 200$ (a pre-registered constant, kept through the relabel below), $K = 8$ channels and $\varepsilon = 0.2$. Trajectories are the cohort's raw discharge capacities. The life label smooths them with a trailing 7-cycle median over full windows (so the reference peak is post-formation; median peak cycle 111), takes the peak of the smoothed curve, and sets L to the first post-peak cycle from which retention stays below 0.8 of that peak for the rest of the record (no such cycle: censored alive at the horizon). A first-crossing rule on retention normalized by the whole-record peak, used in an earlier version, fired on formation transients and labelled 21 of the 36 cells dead at cycles 1–5, 16 of which then held retention above 0.8 for more than 100 further cycles; the rule above gives 34 good cells and 2 short records (140 and 151 cycles) censored alive before the spec, counted not-observed-good as on HUST, so no failure is observed within the record at this spec and recall is reported over those two cells only (0 or 1 loaded per seed). Each seed's campaign loads 13–14 good cells; Rule A stops nothing on 19 of 20 seeds and issues one false stop on the twentieth (worst-seed FSR 0.071, mean 0.004, recall 0); the naive rule's FSR is 0.02 on average and 0.14 on its worst seed, one breach of the budget, at recall 0.15; halving's is 0.53 (0.77 worst) at recall 0.10. Soft rejection (*nosr* ablation rows) is neutral in finite-pool replay, retained in the deployed rule, no claim made on it. Per-seed stop counts in every released result file expose abstention directly.

Baseline configuration. ASHA and the halving stop in BOHB/EI-Hyperband use $\eta = 2$ with per-look quantile retirement (min. 3 observed), single-bracket rank stops, not multi-bracket Hyperband; at $\eta = 2$ the retirement rule is median pruning at the same look. BOHB's tree-structured Parzen estimator (TPE) selector draws uniformly until 6 observations and takes a uniform arm with probability 0.1. Linear extrapolation fits from cycle 10 on at least 10 points and declares end of life at 0.8 of the observed peak capacity, and is regime-bimodal (uselessly conservative on MATR, unsafe on ISU-ILCC at 0.77); the EI surrogate refits every 5 observations and is uniform below 3; the specification-stress rows' forecaster trains on as few as 3 cells where every other row requires 5. The headline false-kill range is not an η artifact. Rerunning all three at $\eta \in \{3, 4\}$ (50 seeds) on the campaign image spreads the family's pooled FSR over 0.46–0.98 across $\eta \in \{2, 3, 4\}$, monotonically *worse* in η, so the shipped $\eta = 2$ rows are the most favorable configuration for the heuristics. BOHB's TPE split uses $\gamma = 0.25$, bandwidth 0.5; the LTT margin baseline uses Clopper–Pearson level $1-\delta = 0.9$ (miscoverage δ, not the halving η above) on the same calibration pool and uniform-random arm selection (its discovery-time rows therefore mix acquisition and stopping effects). Its schedule needs care. A conformal arm may spend α_j at a look once $\alpha_j \geq 1/(n_j + 1)$, but the Clopper–Pearson bound at $k = 0$ is $1 - \delta^{1/n_j}$, which over the deployed n_j (16–35 on the corpora where LTT fires) is 2.28–$2.29\times$ larger. An earlier configuration handed LTT the conformal-feasible uniform split, which its own bound can never clear; it then fired only on the seeds where the conformal filter *itself* failed and fell back to concentrating the budget at the first look, $13/100$ seeds on ISU-ILCC, $5/100$ on MATR, none on HUST. Table 2 therefore schedules LTT by the feasibility routine against its *own* floor, with one further rule: when no multi-look split is feasible the whole budget is placed at the *final* look (the same late-look convention the theorem arms use) rather than at the first, which is where the earlier configuration's fallback landed. (We note the conformal rows additionally receive the SCHED subset-search optimizer; LTT's CP floor admits only single-look schedules here, so the search would reduce to the same look choice.) Attribution differs by cor-

pus. On MATR, whose per-look pools of 16–23 good cells leave every *multi-look split* CP-infeasible (two looks would need $2 \times 0.095 > \alpha$; the final look alone is CP-feasible on the 10 largest-pool seeds), the CP floor alone reproduces the old row exactly and the entire lift $0.001 \to 0.029$ (zero false stops) comes from the final-look placement; on ISU-ILCC the floor itself carries the bulk of $0.026 \to 0.133$ (max seed 0.041), HUST still abstaining. The derived schedule coincides with the campaign's late-look arm on all three corpora, verified at the bit level: the six-corpus extension, the weighted α-sweep, and the budgeted racing/LTT arms were all rerun on the campaign's own cluster image and environment, and the derived-schedule arm's per-seed arrays are identical to the late-look arm's in every field.

How much of the savings is certificate-covered? A first-stop-latched replay of the weighted rule (identical construction, stops disabled after each campaign's first) decomposes its busy-channel savings into the part a first-stop certificate could in principle speak to and the audit-only remainder: the after-first share is 70% on MATR, 93% on ISU-ILCC, and 76% on TONGJI. Most of even the audited rule's savings accrue in the regime no theorem here covers, the open conjecture's territory, and a further reason the negative headline stands. Run on all six corpora the LTT baseline abstains on four. At their $|\mathrm{Cal}_j| \in [6, 20]$ its Clopper–Pearson floor exceeds α outright (it needs 22), so the granularity problem is not specific to conformal calibration. A further fairness arm matches softmax-EI acquisition (100 seeds); matching the acquisition leaves every stop decision unchanged, per-seed false/true stop counts, FSR, and recall are seed-by-seed identical, because the unbudgeted replay tests every cell and each cell's stop is calibration-pool-based (the reported bootstrap intervals can still differ in the last digit: the resampling RNG is keyed by arm), while selection order still reorders the campaign, so discovery times (and which stop is the campaign's first) differ. Both racing and the fairly scheduled LTT were also rerun under the 40% channel-cycle budget, since the unbudgeted replay cannot show a capacity benefit for a policy that only reorders loads. The verdict survives. Racing returns zero recall at exactly no-stop's busy total on all three corpora, and LTT reaches recall 0.030 (MATR, no false stops) and 0.136 (ISU-ILCC, worst seed 0.041) against the weighted rule's 0.055 and 0.259, the same $\sim$0.5–0.6$\times$ ratio as the unbudgeted rows, with busy totals within 0.21% of no-stop for all three of these policies. Racing retires an arm after 1 completed spec-missing replicate with no running-cell stops; fixed-horizon truncation (Attia et al., 2020) stops every cell at cycle 100 and so forfeits every endpoint *inside the loop* by construction, a different objective (fixed-budget best-arm identification, not per-cell certification), so its row quantifies what the paradigm gives up; the tuned-margin baseline thresholds at the per-look empirical α-quantile of the same good-only calibration scores (no conformal correction, no weights, no spending). Core configuration constants are embedded in every released result file; arm-specific flags are carried by the file-name convention and the shipped launch scripts.

Table 6: Generality corpora (100 seeds, $\alpha = 0.1$): FSR mean (max seed) / bad-cell recall; **bold** = worst seed within budget; "—" = no stops fired. SNL: 52 cells, 19 arms ({LFP, NCA, NMC} $\times$ {15, 25, 35}°C $\times$ SOC window $\times$ rates); the weighted gates close almost completely (two stops in 100 seeds, recall 0.002); the pooled unweighted mode clears its floor and fires at recall 0.02 with worst-seed fragility (0.125; $2/100$ violating). TONGJI: 99 cells, 10 arms (3 cathodes $\times$ temperature $\times$ charge rate, plus sub-group splits; 2–26 replicates/arm), calibration sized to two replicates per arm by the deployed feasibility gate's pre-campaign arithmetic (Section 3.2); the weighted rule fires at recall 0.14 with zero worst-seed violations. Every ladder rung replicates (100 seeds; BOHB/EI-Hyperband/extrapolation/fixed-horizon rows: 50); linear extrapolation spans its full regime-bimodality, uselessly conservative on MATR, $\mathrm{FSR} = 1.0$ on TONGJI.

Policy	SNL	TONGJI
ASHA (halving)	0.749 (1.000) / 0.42	0.371 (0.714) / 0.90
BOHB	0.761 (1.000) / 0.44	0.360 (0.682) / 0.88
EI-Hyperband	0.756 (1.000) / 0.45	0.429 (0.810) / 0.92
Linear extrapolation	0.526 (0.700) / 0.95	1.000 (1.000) / 1.00
Fixed-horizon (cycle 100)	1.000 (1.000) / 1.00	1.000 (1.000) / 1.00
Naive $\hat{\mu} < L_{\mathrm{spec}}$	0.311 (1.000) / 0.72	0.265 (0.840) / 0.98
Naive + tuned margin	0.040 (0.444) / 0.19	0.034 (0.391) / 0.87
LTT margin (CP-UCB, own floor)	**0.000 (0.000) / —**	**0.000 (0.000) / —**
Good-only conformal, unweighted	0.004 (0.125) / 0.02	0.010 (0.174) / 0.68
+ weighted FCS (ours)	**0.000 (0.000) / 0.00**	**0.000 (0.045) / 0.14**

Beyond batteries (Table 7). One cross-domain corpus tests whether the mechanism is battery-specific. IVPT is a *real* in-vitro permeation-testing corpus (135 Franz-cell runs, 28 topical formulation arms over P407 $\times$ ethanol $\times$ PG, 3–11 replicates; SHA-256 provenance ships with the code). The same 135-run corpus is the permeation dataset of Xiao et al. (2026), there used for threshold classification rather than irreversible hitting-time screening. Goodness (reaching the median permeation target within 30 h) is a hitting time of an *increasing* process, the mirror of battery survival; the loader recodes it monotonically as a pseudo-life ($10 \times (31 - \text{crossing hour})$; never-crossed $\mapsto$ below spec), after which every mechanism applies unchanged. The looks-precede-goodness requirement holds on the time axis for 133 of 135 runs (decisions at $\{4, 8, 22\}$ h; 63 of the 68 crossings fall in 24–30 h). Two runs cross at 21 and 22 h (no later than the final look) so unlike the battery corpora the requirement is not asserted at runtime here; we report the arm as cross-domain evidence with that exception stated. One convention artifact was found and fixed. The corpus samples at hours $\{1, 2, 3, 4, 6, 8, 22, 24, 26, 28, 30\}$, so an hourly interpolant gave an earlier $\{12, 18, 22\}$-look convention features anchored at the 22 h sample, information not causally available at those hours. The canonical rows use the causal looks $\{4, 8, 22\}$ h, whose interpolated prefixes touch only raw samples at or before each look (the short-prefix feature anchor this needs is bit-identical for the battery corpora): weighted recall 0.20, max seed 0.097, zero violating seeds. Under the noncausal convention the weighted mode read 0.245 with one breach (0.129); those files are retained for comparison. The 22 h look and all goodness labels were never affected, and the battery corpora have no such gap. The crossing hour, and with it the pseudo-life label, is read on the hourly linear interpolant of the 11 raw samples (29 of the 68 crossings fall at non-sample hours); never-crossed runs carry the constant pseudo-life 5 (67 of 135), and $L_{\text{spec}} = 10$ (the pool median) corresponds to crossing within 30 h. Calibration is sized by the same pre-campaign granularity arithmetic as TONGJI, which here yields one reserved replicate per arm.

Table 7: Beyond batteries (100 seeds, $\alpha = 0.1$): FSR mean (max seed) / recall / violating seeds; **bold** = worst seed within budget, as in Table 2. IVPT: a *real* permeation-screening corpus of 135 Franz-cell runs over 28 topical formulation arms (P407 $\times$ ethanol $\times$ PG; 3–11 replicates), goodness = reaching the permeation target, recoded as a pseudo-life hitting time (Appendix note); decisions at the causal looks $\{4, 8, 22\}$ h (raw sampling hours) precede the 24–30 h crossing region (an earlier noncausal $\{12, 18, 22\}$ convention is disclosed in the corpus note). The ladder replicates end-to-end; the weighted rule's development-seed worst seeds sit within budget (IVPT 0.097), though its held-out bank shows a $1/50$ breach on this corpus under each look convention (Appendix C), consistent with its audit-only status.

Policy	IVPT (real)
ASHA (halving)	0.560 (0.808) / 0.96 / 100/100
Naive $\hat{\mu} <$ spec	0.332 (0.618) / 0.99 / 100/100
Naive + tuned margin	0.104 (0.324) / 0.76 / 46/100
LTT margin (CP-UCB, own floor)	**0.000 (0.000) / — / 0/100**
Good-only conformal, unweighted	0.037 (0.270) / 0.43 / 7/100
+ weighted FCS (ours)	**0.018 (0.097) / 0.20 / 0/100**

Table 8: Complete results: every (experiment, policy) of the canonical sunk-convention campaign with seed-block bootstrap 95% CIs; FSR max is the worst seed; Viol. counts seeds whose per-seed FSR exceeds that experiment's budget α.

Experiment	Policy	Seeds	FSR mean [CI]	FSR max	Viol.	Recall mean [CI]
hust_a0.05	rule_a	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
hust_a0.1	rule_a	50	0.012 [0.005, 0.021]	0.154	1/50	0.085 [0.056, 0.116]
hust_a0.2	rule_a	50	0.050 [0.031, 0.071]	0.308	3/50	0.204 [0.153, 0.257]
hust_ablations	rule_a_nosr	50	0.012 [0.005, 0.021]	0.154	1/50	0.085 [0.057, 0.116]
hust_ablations	rule_a_uni	50	0.012 [0.005, 0.021]	0.154	1/50	0.066 [0.044, 0.091]
hust_ablations	rule_a_mixedcal	50	0.031 [0.017, 0.047]	0.235	6/50	0.135 [0.098, 0.175]
hust_baselines	bohb	50	0.487 [0.438, 0.532]	0.750	49/50	0.828 [0.802, 0.854]
hust_baselines	hyperband_ei	50	0.452 [0.409, 0.496]	0.812	50/50	0.808 [0.776, 0.838]
hust_baselines	linextrap	50	0.355 [0.326, 0.384]	0.667	50/50	0.542 [0.516, 0.567]
hust_baselines	fixed100	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
hust_baselines	rule_b	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
hust_eps005	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_eps01	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_eps03	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_eps05	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_litspec	rule_a	100	0.009 [0.004, 0.014]	0.125	2/100	0.070 [0.051, 0.089]
hust_litspec	rule_a_w	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_litspec_t2	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_main	rule_a	100	0.015 [0.007, 0.023]	0.267	4/100	0.089 [0.068, 0.112]
hust_main	naive	100	0.549 [0.494, 0.602]	1.000	92/100	0.756 [0.710, 0.799]
hust_main	asha	100	0.475 [0.442, 0.508]	0.750	98/100	0.805 [0.785, 0.826]
hust_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_mainplus	naive_margin	100	0.064 [0.048, 0.082]	0.467	24/100	0.256 [0.221, 0.292]
hust_mainw	rule_a_w	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_panel2	ltt_margin	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_panel3	ltt_margin_late	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_panel3	ltt_margin_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_panel3	ltt_margin_ei_late	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_theorem2	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_theorem2	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_theorem	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_theorem	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
hust_wsplit	rule_a_w_split	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_a0.05	rule_a	50	0.007 [0.003, 0.010]	0.055	1/50	0.180 [0.131, 0.234]
isu_a0.1	rule_a	50	0.019 [0.013, 0.026]	0.111	1/50	0.381 [0.327, 0.437]
isu_a0.2	rule_a	50	0.028 [0.021, 0.035]	0.111	0/50	0.531 [0.481, 0.580]
isu_ablations	rule_a_nosr	50	0.019 [0.013, 0.026]	0.111	1/50	0.381 [0.327, 0.437]
isu_ablations	rule_a_uni	50	0.010 [0.005, 0.015]	0.111	1/50	0.253 [0.194, 0.316]
isu_ablations	rule_a_mixedcal	50	0.022 [0.016, 0.030]	0.130	1/50	0.447 [0.402, 0.493]
isu_baselines	bohb	50	0.954 [0.942, 0.964]	1.000	50/50	0.572 [0.526, 0.617]
isu_baselines	hyperband_ei	50	0.947 [0.936, 0.958]	1.000	50/50	0.578 [0.540, 0.617]
isu_baselines	linextrap	50	0.768 [0.758, 0.778]	0.863	50/50	0.988 [0.984, 0.991]
isu_baselines	fixed100	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
isu_baselines	rule_b	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
isu_degfc	rule_a	50	0.021 [0.014, 0.028]	0.111	1/50	0.375 [0.317, 0.432]
isu_degfc	rule_a_w	50	0.011 [0.007, 0.015]	0.053	0/50	0.218 [0.186, 0.251]
isu_eps005	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_eps01	rule_a_online	100	0.001 [0.000, 0.002]	0.066	0/100	0.002 [0.000, 0.004]
isu_eps03	rule_a_online	100	0.002 [0.001, 0.004]	0.043	0/100	0.010 [0.005, 0.015]
isu_eps05	rule_a_online	100	0.001 [0.000, 0.001]	0.018	0/100	0.002 [0.000, 0.004]
isu_litspec	rule_a	100	0.017 [0.013, 0.022]	0.125	1/100	0.338 [0.304, 0.373]
isu_litspec	rule_a_w	100	0.010 [0.007, 0.013]	0.065	0/100	0.210 [0.187, 0.234]
isu_litspec_t2	rule_a_online	100	0.007 [0.004, 0.011]	0.114	1/100	0.029 [0.020, 0.039]
isu_main	rule_a	100	0.016 [0.012, 0.021]	0.111	1/100	0.388 [0.351, 0.426]
isu_main	naive	100	0.180 [0.166, 0.194]	0.500	89/100	0.950 [0.943, 0.956]
isu_main	asha	100	0.957 [0.951, 0.964]	1.000	100/100	0.587 [0.561, 0.613]

Table 9: Complete results (continued): every (experiment, policy) of the canonical sunk-convention campaign with seed-block bootstrap 95% CIs; FSR max is the worst seed; Viol. counts seeds whose per-seed FSR exceeds that experiment's budget α.

Experiment	Policy	Seeds	FSR mean [CI]	FSR max	Viol.	Recall mean [CI]
isu_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_mainplus	naive_margin	100	0.029 [0.023, 0.035]	0.130	3/100	0.580 [0.542, 0.617]
isu_mainw2	rule_a_w	100	0.008 [0.005, 0.011]	0.077	0/100	0.188 [0.164, 0.212]
isu_mainw	rule_a_w	100	0.008 [0.006, 0.011]	0.050	0/100	0.225 [0.200, 0.250]
isu_nodqv	rule_a	50	0.017 [0.011, 0.025]	0.111	1/50	0.384 [0.328, 0.440]
isu_nodqv	rule_a_w	50	0.009 [0.005, 0.013]	0.053	0/50	0.225 [0.189, 0.262]
isu_panel2	ltt_margin	100	0.001 [0.000, 0.002]	0.033	0/100	0.026 [0.010, 0.044]
isu_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_panel3	ltt_margin_late	100	0.003 [0.001, 0.005]	0.041	0/100	0.133 [0.104, 0.164]
isu_panel3	ltt_margin_ei	100	0.001 [0.000, 0.002]	0.033	0/100	0.026 [0.010, 0.045]
isu_panel3	ltt_margin_ei_late	100	0.003 [0.001, 0.005]	0.041	0/100	0.133 [0.104, 0.164]
isu_theorem2	rule_a_online	100	0.007 [0.003, 0.012]	0.158	1/100	0.030 [0.020, 0.042]
isu_theorem2	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_theorem	rule_a_online	100	0.001 [0.000, 0.003]	0.041	0/100	0.008 [0.004, 0.013]
isu_theorem	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isu_wsplit	rule_a_w_split	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isuefc_main	rule_a	100	0.016 [0.012, 0.020]	0.122	1/100	0.073 [0.059, 0.087]
isuefc_main	naive	100	0.495 [0.463, 0.527]	1.000	100/100	0.838 [0.815, 0.858]
isuefc_main	asha	100	0.839 [0.825, 0.853]	0.974	100/100	0.672 [0.653, 0.690]
isuefc_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isuefc_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
isuefc_mainw	rule_a_w	100	0.009 [0.006, 0.012]	0.077	0/100	0.052 [0.042, 0.063]
ivpt_main	rule_a	100	0.032 [0.024, 0.041]	0.171	9/100	0.501 [0.443, 0.559]
ivpt_main	naive	100	0.119 [0.107, 0.132]	0.310	55/100	0.981 [0.974, 0.987]
ivpt_main	asha	100	0.535 [0.511, 0.559]	0.808	100/100	0.956 [0.949, 0.964]
ivpt_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivpt_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivpt_mainplus	naive_margin	100	0.065 [0.055, 0.074]	0.200	23/100	0.780 [0.736, 0.823]
ivpt_mainw	rule_a_w	100	0.016 [0.011, 0.022]	0.129	1/100	0.245 [0.215, 0.275]
ivpt_panel2	ltt_margin	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivpt_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivptc_main	rule_a	100	0.037 [0.029, 0.047]	0.270	7/100	0.426 [0.364, 0.488]
ivptc_main	naive	100	0.332 [0.311, 0.352]	0.618	100/100	0.991 [0.986, 0.995]
ivptc_main	asha	100	0.560 [0.538, 0.583]	0.808	100/100	0.957 [0.948, 0.964]
ivptc_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivptc_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivptc_mainplus	naive_margin	100	0.104 [0.090, 0.119]	0.324	46/100	0.759 [0.715, 0.802]
ivptc_mainw	rule_a_w	100	0.018 [0.013, 0.023]	0.097	0/100	0.201 [0.169, 0.233]
ivptc_panel2	ltt_margin	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
ivptc_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_a0.05	rule_a	50	0.005 [0.002, 0.008]	0.057	1/50	0.323 [0.258, 0.388]
matr_a0.1	rule_a	50	0.012 [0.007, 0.019]	0.105	1/50	0.529 [0.477, 0.577]
matr_a0.2	rule_a	50	0.025 [0.017, 0.034]	0.162	0/50	0.657 [0.611, 0.699]
matr_ablations	rule_a_nosr	50	0.012 [0.007, 0.019]	0.105	1/50	0.529 [0.478, 0.577]
matr_ablations	rule_a_uni	50	0.009 [0.003, 0.016]	0.114	2/50	0.032 [0.023, 0.042]
matr_ablations	rule_a_mixedcal	50	0.016 [0.009, 0.025]	0.162	2/50	0.557 [0.507, 0.605]
matr_baselines	bohb	50	0.819 [0.805, 0.833]	0.919	50/50	0.806 [0.784, 0.827]
matr_baselines	hyperband_ei	50	0.822 [0.807, 0.836]	0.909	50/50	0.803 [0.782, 0.823]
matr_baselines	linextrap	50	0.000 [0.000, 0.000]	0.000	0/50	0.057 [0.051, 0.064]
matr_baselines	fixed100	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
matr_baselines	rule_b	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
matr_degfc	rule_a	50	0.013 [0.006, 0.021]	0.154	1/50	0.187 [0.144, 0.233]
matr_degfc	rule_a_w	50	0.004 [0.002, 0.008]	0.061	0/50	0.018 [0.009, 0.028]
matr_eps005	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_eps01	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_eps03	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_eps05	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.001]

Table 10: Complete results (continued): every (experiment, policy) of the canonical sunk-convention campaign with seed-block bootstrap 95% CIs; FSR max is the worst seed; Viol. counts seeds whose per-seed FSR exceeds that experiment's budget α.

Experiment	Policy	Seeds	FSR mean [CI]	FSR max	Viol.	Recall mean [CI]
matr_holdb1	rule_a	50	0.032 [0.016, 0.052]	0.304	4/50	0.292 [0.214, 0.377]
matr_holdb1	rule_a_w	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
matr_holdb2	rule_a	50	0.000 [0.000, 0.000]	0.000	0/50	0.034 [0.003, 0.078]
matr_holdb2	rule_a_w	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
matr_holdb3	rule_a	50	0.022 [0.016, 0.029]	0.108	1/50	0.300 [0.220, 0.387]
matr_holdb3	rule_a_w	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
matr_litspec	rule_a	100	0.011 [0.008, 0.016]	0.140	1/100	0.560 [0.524, 0.595]
matr_litspec	rule_a_w	100	0.004 [0.002, 0.007]	0.093	0/100	0.051 [0.041, 0.062]
matr_litspec_t2	rule_a_online	100	0.001 [0.000, 0.003]	0.062	0/100	0.004 [0.002, 0.007]
matr_main	rule_a	100	0.012 [0.008, 0.017]	0.167	2/100	0.536 [0.499, 0.573]
matr_main	naive	100	0.475 [0.450, 0.500]	0.846	100/100	0.959 [0.953, 0.966]
matr_main	asha	100	0.818 [0.807, 0.829]	0.927	100/100	0.805 [0.788, 0.821]
matr_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_mainplus	naive_margin	100	0.034 [0.026, 0.043]	0.262	6/100	0.682 [0.650, 0.712]
matr_mainw2	rule_a_w	100	0.004 [0.002, 0.006]	0.054	0/100	0.048 [0.038, 0.058]
matr_mainw	rule_a_w	100	0.005 [0.002, 0.008]	0.095	0/100	0.051 [0.041, 0.062]
matr_nodqv	rule_a	50	0.010 [0.004, 0.017]	0.114	2/50	0.045 [0.031, 0.061]
matr_nodqv	rule_a_w	50	0.005 [0.002, 0.008]	0.057	0/50	0.004 [0.002, 0.007]
matr_panel2	ltt_margin	100	0.001 [0.000, 0.003]	0.061	0/100	0.001 [0.000, 0.002]
matr_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_panel3	ltt_margin_late	100	0.000 [0.000, 0.000]	0.000	0/100	0.029 [0.011, 0.051]
matr_panel3	ltt_margin_ei	100	0.001 [0.000, 0.003]	0.061	0/100	0.001 [0.000, 0.002]
matr_panel3	ltt_margin_ei_late	100	0.000 [0.000, 0.000]	0.000	0/100	0.029 [0.011, 0.051]
matr_theorem2	rule_a_online	100	0.001 [0.000, 0.003]	0.062	0/100	0.003 [0.001, 0.004]
matr_theorem2	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_theorem	rule_a_online	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.001]
matr_theorem	rule_a_block	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
matr_wsplit	rule_a_w_split	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
snl_baselines	bohb	50	0.761 [0.711, 0.806]	1.000	50/50	0.436 [0.387, 0.486]
snl_baselines	hyperband_ei	50	0.756 [0.705, 0.808]	1.000	50/50	0.447 [0.403, 0.492]
snl_baselines	linextrap	50	0.526 [0.504, 0.547]	0.700	50/50	0.952 [0.937, 0.966]
snl_baselines	fixed100	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
snl_baselines	rule_b	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
snl_main	rule_a	100	0.004 [0.001, 0.009]	0.125	2/100	0.020 [0.007, 0.036]
snl_main	naive	100	0.311 [0.270, 0.354]	1.000	83/100	0.724 [0.671, 0.774]
snl_main	asha	100	0.749 [0.716, 0.781]	1.000	100/100	0.415 [0.385, 0.447]
snl_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
snl_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
snl_mainplus	naive_margin	100	0.040 [0.027, 0.056]	0.444	9/100	0.186 [0.148, 0.226]
snl_mainw	rule_a_w	100	0.000 [0.000, 0.000]	0.000	0/100	0.002 [0.000, 0.005]
snl_panel2	ltt_margin	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
snl_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
tongji_baselines	bohb	50	0.360 [0.322, 0.401]	0.682	49/50	0.884 [0.864, 0.905]
tongji_baselines	hyperband_ei	50	0.429 [0.381, 0.479]	0.810	47/50	0.915 [0.897, 0.933]
tongji_baselines	linextrap	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
tongji_baselines	fixed100	50	1.000 [1.000, 1.000]	1.000	50/50	1.000 [1.000, 1.000]
tongji_baselines	rule_b	50	0.000 [0.000, 0.000]	0.000	0/50	0.000 [0.000, 0.000]
tongji_main	rule_a	100	0.010 [0.006, 0.016]	0.174	2/100	0.681 [0.634, 0.725]
tongji_main	naive	100	0.265 [0.237, 0.296]	0.840	88/100	0.980 [0.967, 0.990]
tongji_main	asha	100	0.371 [0.343, 0.400]	0.714	97/100	0.896 [0.881, 0.910]
tongji_main	no_stop	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
tongji_main	no_stop_ei	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
tongji_mainplus	naive_margin	100	0.034 [0.023, 0.048]	0.391	9/100	0.869 [0.841, 0.895]
tongji_mainw	rule_a_w	100	0.000 [0.000, 0.001]	0.045	0/100	0.143 [0.123, 0.165]
tongji_panel2	ltt_margin	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]
tongji_panel2	racing	100	0.000 [0.000, 0.000]	0.000	0/100	0.000 [0.000, 0.000]

Table 11: Operational cost, 100 seeds, completions-only convention: cycles until the first top-decile arm *completes* testing (an early stop reveals no endpoint and is not a discovery; mean [bootstrap 95% CI]; "$(k/100)$" marks seeds that observed any top-decile endpoint at all; means over fewer seeds are conditional). Under this metric, aggressive stopping's apparent discovery speed evaporates. ASHA's ISU-ILCC figure averages only the seeds in which it observed a winner at all, and those are the easy seeds. Matched on the same seeds it is significantly *slower* than no-stop ($+1918$ cycles $[937, 3147]$ over 24 matched seeds). On HUST it does gain through channel freeing ($1.11\times$ faster, $[-482, -141]$), and on MATR the paired delta is null ($+17$ $[-78, 112]$); the unpaired columns overstate all three because exit-based metrics credited killed winners. The static and weighted modes match the acquisition-matched EI no-stop and are faster on ISU-ILCC (-921 and -935 cycles, seed-paired); the online shared mode is slower by 246 (MATR) and 500 (HUST) cycles, the same purchased ε-stream occupancy its busy premium shows. Bottom block: busy channel-kilocycles consumed; savings scale with recall (static 15%/5%/4%; weighted 2%/3%/0%). The two corpora not shown in the columns above give weighted savings SNL 0.0%, TONGJI 3.1%; Section 4's $\sim$3% ceiling is the maximum over all five. The online shared mode's busy *premium* over no-stop is its purchased ε-stream occupancy, charged in full. Racing never stops a running cell; its retirements only *reorder* loads, so its busy cycles coincide with no-stop exactly. That is not an artifact of the unbudgeted convention. Rerun under the 40% budget it still returns zero recall at no-stop's busy total on all three corpora (App. C).

	MATR	ISU-ILCC	HUST
Cycles to first top-decile completion			
No-stop (uniform loading)	1344.5 [1242.9, 1449.9]	13641.6 [12814.0, 14494.8]	3484.1 [3229.7, 3753.7] (99/100)
No-stop (matched EI acquisition)	1479.1 [1361.8, 1606.5]	13802.4 [12902.0, 14782.3]	3595.1 [3341.3, 3855.4] (99/100)
ASHA (halving)	1346.4 [1260.9, 1440.5] (99/100)	13383.7 [12292.3, 14457.4] (24/100)	2851.9 [2729.3, 2992.5] (78/100)
Naive $\hat{\mu} < L_{\text{spec}}$	1295.1 [1227.3, 1368.9]	12313.8 [11537.2, 13125.0]	3105.5 [2904.7, 3325.2] (73/100)
Naive + tuned margin	1288.1 [1213.5, 1365.7]	13282.2 [12431.2, 14195.5]	3443.1 [3207.8, 3692.2] (97/100)
LTT margin (CP-UCB, own floor)	1340.4 [1238.6, 1444.9]	13548.3 [12723.9, 14416.0]	3484.1 [3231.1, 3755.9] (99/100)
Racing (arm-level)	1366.0 [1263.8, 1477.8]	13115.9 [12384.6, 13853.4]	3484.1 [3235.7, 3754.5] (99/100)
Good-only conformal (static)	1363.8 [1270.8, 1463.4]	12881.3 [12138.4, 13682.2]	3594.2 [3341.3, 3849.8] (99/100)
+ weighted FCS (ours)	1515.0 [1398.4, 1636.2]	12867.1 [12101.4, 13678.8]	3595.1 [3342.5, 3851.3] (99/100)
Online shared	1725.0 [1562.3, 1894.8]	13892.1 [12963.2, 14845.7]	4095.4 [3716.5, 4489.6] (99/100)
Busy channel-kilocycles			
No-stop (uniform loading)	65.8 [65.5, 66.2]	534.4 [528.6, 540.4]	56.8 [56.4, 57.1]
No-stop (matched EI acquisition)	65.8 [65.5, 66.2]	534.4 [528.5, 540.2]	56.8 [56.4, 57.1]
ASHA (halving)	17.6 [17.0, 18.3]	41.4 [37.5, 45.6]	23.5 [22.2, 24.9]
Naive $\hat{\mu} < L_{\text{spec}}$	27.2 [26.1, 28.2]	386.2 [376.7, 395.4]	22.5 [20.0, 25.0]
Naive + tuned margin	52.3 [51.5, 53.1]	487.5 [479.7, 495.3]	49.4 [48.3, 50.5]
LTT margin (CP-UCB, own floor)	65.4 [64.9, 65.9]	527.2 [521.1, 533.5]	56.8 [56.4, 57.1]
Racing (arm-level)	65.8 [65.5, 66.2]	534.4 [528.7, 540.2]	56.8 [56.4, 57.1]
Good-only conformal (static)	56.1 [55.3, 56.9]	507.9 [500.7, 515.0]	54.5 [53.8, 55.1]
+ weighted FCS (ours)	64.6 [64.2, 65.1]	518.3 [512.0, 524.7]	56.8 [56.4, 57.1]
Online shared	71.0 [70.3, 71.7]	540.0 [534.0, 546.0]	67.3 [66.3, 68.3]

Table 12: Budgeted regime (100 seeds; campaign truncated when cumulative channel-cycle *consumption* reaches 40% of the pool total; mean [bootstrap 95% CI]). Top: cycles until the first top-decile arm *completes* (endpoint observed; "$(k/100)$" marks how many seeds observed any top-decile endpoint at all). Bottom: observed-endpoint regret. The EI-matched no-stop row isolates acquisition from stopping. On ISU-ILCC the regret gain over uniform no-stop comes from EI acquisition (on MATR it is within noise), and calibrated stopping adds *safety at no regret cost*; seed-paired deltas versus EI no-stop are null on ISU-ILCC ($+1$ $[-7, 9]$ static, $+4$ $[-5, 16]$ weighted), -0.1 static and exactly zero for the abstaining weighted rule on HUST, and favorable but unresolved on MATR (-34 $[-77, 1]$, -24 $[-55, 1]$). ASHA destroys the campaign's product. It observes *no* top-decile endpoint in $76/100$ ISU-ILCC seeds and its observed regret explodes. Under truncation the weighted rule's worst-seed FSR is $0.108/0.058/0.000$ over cells resolved before the cut (the loaded-cell denominator puts these at 0.089 and 0.050 or above, App. C); one MATR seed exceeds α $(1/100$; the unbudgeted campaign's worst seed is $0.095)$. Tested-arm regret is blind to all of this. In unbudgeted campaigns the static and weighted modes match EI no-stop on discovery time and are earlier on ISU-ILCC (seed-paired -921 $[-1675, -237]$ static, -935 $[-1679, -260]$ weighted; Table 11).

	MATR	ISU-ILCC	HUST
Cycles to first top-decile completion			
No-stop (uniform loading)	1345 [1243, 1450]	13642 [12799, 14497]	3484 [3226, 3757] (99/100)
No-stop (matched EI acquisition)	1479 [1358, 1604]	13802 [12888, 14766]	3595 [3346, 3854] (99/100)
ASHA (halving)	1346 [1259, 1438] (99/100)	13384 [12284, 14475] (24/100)	2852 [2730, 2989] (78/100)
Good-only conformal (static)	1364 [1267, 1465]	12881 [12136, 13656]	3594 [3343, 3855] (99/100)
+ weighted FCS (ours)	1515 [1398, 1633]	12867 [12102, 13648]	3595 [3348, 3851] (99/100)
Observed-endpoint regret (cycles)			
No-stop (uniform loading)	162 [99, 230]	40 [19, 63]	41 [28, 55]
No-stop (matched EI acquisition)	123 [70, 181]	18 [3, 36]	41 [28, 55]
ASHA (halving)	175 [113, 242]	2143 [1946, 2332]	119 [94, 146]
Good-only conformal (static)	90 [45, 138]	19 [6, 35]	41 [28, 55]
+ weighted FCS (ours)	99 [53, 152]	22 [6, 38]	41 [28, 55]