DEPLOYMENT RISK CONTROL FOR SEQUENTIAL PREDICTION SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployed prediction systems are maintained while they serve: operators recalibrate thresholds, repair inputs, switch to backups, or abstain, often on the system's own feedback. Yet the available controllers certify pieces of this loop—a coverage recursion, a change detector, a single gate—and no unit of accountability spans the interventions, the outputs they serve, and the decisions they drive. We show that deployment risk control is invariance control of one scalar: the running risk-budget balance. Formalizing deployment as a *sequential intervention process* (SIP), we prove a characterization: a controller is budget-compliant on every trajectory *iff* it confines this scalar to an invariant interval, *iff* exponential transforms of the balance are universal-null e-processes; at the boundaries compliance forces the controller's action, so all design freedom lives strictly inside. A constructive two-ledger theorem achieves the invariant: a deterministic ledger keeps the *booked risk proxy* within $O(1/\gamma T)$ of budget on every trajectory, and an intervention ledger bounds interventions fired while their nulls hold, uniformly over stopping times. A certified bridge carries the same ledgers to decision losses, and a gated meta-policy selects among the controllers themselves under one composite anytime-valid budget. We recover ACI, conformal PID, and online CRC as special cases. Under injected degradation on six real streams, the unified policy holds the booked budget at 2.4–2.7$\times$ smaller width than NexCP, the only one of three external baselines that also holds it there; replaying the CDC FluSight and COVID-19 hubs end to end, the certified answer was restraint.

1 INTRODUCTION

Modern sequential prediction deployments (forecasting services, monitored regression endpoints, clinical early-warning scores, threshold-served content filters) are closed-loop systems. When a monitored electricity-price model starts missing its targets, an operator does not watch passively: they widen prediction intervals, re-estimate the calibration quantile, repair corrupted inputs, switch to a robust baseline, or take the model offline. Each intervention changes the loss sequence that later interventions observe: the policy and the data it intervenes on are entangled.

This problem is not specific to forecasting. We call it *sequential risk-budgeted intervention*: given a served pipeline, a library of interventions, and delayed feedback, spend history-dependent interventions against explicit risk budgets so that the delivered trajectory, the one the user actually received, retains guarantees at every time and every stopping time. The formalism itself assumes nothing beyond these primitives plus a bounded per-step risk proxy (bounded losses for the decision layer); the theorems' remaining premises (boundary feasibility, gate nulls, meta-layer exogeneity) are stated where used. Time-series forecasting, where deployment-time intervention is most institutionalized, serves as the primary instantiation and testbed; we also report a non-forecasting instantiation (online classification risk control).

The community has built rigorous tools for individual arcs of this loop: adaptive conformal inference (ACI; Gibbs & Candès, 2021) recalibrates a served quantile online with long-run coverage under arbitrary shift; conformal PID control (Angelopoulos et al., 2023) adds integrator and scorecaster terms; conformal risk control (CRC; Angelopoulos et al., 2024; Feldman et al., 2023) generalizes coverage to any bounded monotone risk. On the structural side, e-processes (Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024) yield anytime-valid certificates for switching decisions. Recent

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.

deployment controllers additionally use debt accounting for stratified recalibration-plus-repair, no-regret operator orchestration, and e-gated model fallback.[1]

What is missing, however, is their composition. A practitioner who runs ACI *and* a fallback rule *and* an input-repair trigger has no guarantee covering the composite: ACI's ledger controls the delivered miscoverage it sees but is silent on whether interventions were warranted; the fallback certificate assumes no concurrent recalibration; and the repair trigger is typically ungated. Nor is the failure hypothetical: a recursion that books a proxy of the pre-intervention output rather than the delivered one can hold delivered miscoverage a constant above budget forever.

This paper. Our central claim is that *deployment risk control is invariance control of one scalar, the running risk-budget balance*. We show that every controller compliant on every trajectory confines this scalar to an invariant interval; at the boundaries, compliance forces its action, so all design freedom lives strictly inside. In this view, structural interventions are discrete actuators whose licenses are bought with e-process wealth. We formalize deployment as a *sequential intervention process* (SIP; Section 2): a policy observes prediction-time evidence, selects an intervention a_t from a library $\mathcal{A}$, and the pipeline emits an output. A bounded *risk proxy* of the delivered output (an observable score, not the unobserved true harm) is charged to a running budget. The key structural insight is the library's dichotomy: *calibration actuators* $\mathcal{A}_{\mathrm{C}}$ act on a scalar served threshold; *structural actuators* $\mathcal{A}_{\mathrm{S}}$ (repair, switch, fallback, abstain) change the pipeline itself, hence can perturb the future loss sequence arbitrarily (Figure 1, left). Three axioms then discipline the loop: (A1) the served threshold follows a budget recursion; (A2) every step is booked exactly once, against the output the user actually received; (A3) every structural state change is gated by an e-process with an α_{S} spending schedule.

Under these axioms we prove a *characterization*: budget compliance $\equiv$ confinement of a single scalar to an invariant interval $\equiv$ a universal-null e-process property—delimiting the design space before any controller is chosen. A constructive *two-ledger* theorem then achieves the invariant for the full library, and the two ledgers are coupled: the gate pins the deterministic ledger's constant to the environment's regime changes rather than the policy's activity.

Empirically, we run nine policies—static split conformal, the five recovered special cases that stand alone as policies (the sixth, operator orchestration, is the gate every gated run exercises), the two strongest published calibration-only baselines (DtACI, betting-ACI), and the unified policy—under one shared budget $(\alpha, \alpha_{\mathrm{S}}) = (0.1, 0.05)$ on six real streams, three scenarios, ten seeds each, plus head-to-head comparisons against three *independent external baselines*. Under degradation, compliance alone does not separate the field (best recalibration-only shortfall 0.005 vs 0.006); the price does: the unified policy holds the budget at 2.4–$2.7\times$ smaller width than conformal PI *and* 2.2–$2.6\times$ smaller than DtACI and betting-ACI, with lower graded risk. The width gap is a decision-loss gap on two operational tasks (newsvendor, storage arbitrage). On real deployments, two natural regime breaks are absorbed by the deterministic ledger while the e-gates correctly stay silent, and two recorded feed-sensor outages supply the complementary power evidence (firing 134 h and 44 h after the feeds die). We also chart where the library costs rather than pays (clean-stream repair cost, dead weight on natural breaks, collapse under dual-view corruption).

The contributions sit at three layers, each defensible on its own. (i) *Theory: the characterization* (Section 3.1): the equivalence stated above, proved with explicit constant translations (Theorem A); (ii) *method: the machinery that achieves the invariant* (Sections 3 and 4): the SIP formalism and axioms A1–A3; the two-ledger theorem (Theorem 3.1), its coupled refinement (Theorem 3.3), and failure boundaries showing each axiom necessary (Section 3.5); a decision layer bridging the ledgers to decision loss (Theorem 4.2), recovering the prediction-set decision and robustness-control objectives (Wang & Dobriban, 2026; Hu et al., 2026) as static special cases and conformal policy control (Prinster et al., 2026) up to structural identification; and a gated meta-policy layer with matching lower bounds, closing on the composite served-trajectory ledger (Appendix G). We recover ACI, conformal PID, and online CRC as projections (Section 3.4; Figure 1, right), so this layer stands on its constructive guarantees alone; (iii) *empirical evidence, released as a reusable diagnostic suite* (Section 5): a cross-framework protocol under one shared budget on six real streams and four backbone families, against independent external baselines (SPCI, NexCP, drift-detection recalibration), two operational decision

[1]Three anonymized third-party manuscripts under review, unverifiable in print under blind review; Appendix D restates their target guarantees verbatim and proves the recoveries self-contained; the supplementary material ships executable reimplementations (`debt_repair`, `efallback`; orchestration is the singleton A3 gate every gated run exercises), so no main-text claim depends on their content.

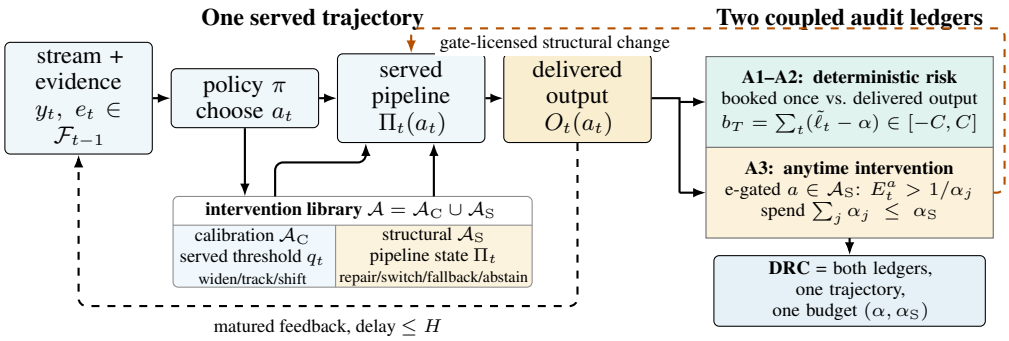

Figure 1: **Deployment risk control makes one served trajectory auditable.** *Left* (SIP loop): evidence drives the policy's choice a_t from a dichotomous library—calibration actuators $\mathcal{A}_C$ tune the served threshold q_t, structural actuators $\mathcal{A}_S$ (repair, switch, fallback, abstain) move the pipeline state Π_t—and matured feedback (delay $\leq H$) closes the loop. *Right*: the trajectory is booked on two coupled ledgers—A1–A2 record each delivered risk proxy once against α, confining the balance b_T; A3 admits a structural change only when its e-process clears the spending gate, licensing it back into the pipeline at total spend α_S. DRC is the composition of the two ledgers, not unrelated controllers.

tasks, and two forecast-hub replays, with a full account of where the library is dead weight. All raw outputs and analysis scripts are released: every reported number is independently recomputable.

1.1 Related work

We relate six strands, each identified by the layer of the SIP stack it occupies; Appendix A discusses each in full. *Online conformal recalibration* (Gibbs & Candès, 2021; Zaffran et al., 2022; Bhatnagar et al., 2023; Angelopoulos et al., 2023; Gibbs & Candès, 2024; Podkopaev et al., 2024) and change-point-aware variants (Sun & Yu, 2025): Theorem 3.1(a) is the common generalization to arbitrary structural interventions under single booking; such variants adapt the served set itself, orthogonal to the certified structural changes of A3. *Risk control beyond coverage* (Angelopoulos et al., 2024; Feldman et al., 2023; Opoku & Banahene, 2026; Kotte, 2026): Theorem 3.1(a) holds for any bounded risk proxy; online CRC is recovered. *Anytime-valid sequential testing* (Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024; Podkopaev & Ramdas, 2022; Shin et al., 2024; Vovk et al., 2021): A3 lifts the single-null construction to a family of nulls with shared spending. *Sufficiency vs. necessity* (Rakhlin & Sridharan, 2017; Vovk et al., 2005; Ruf et al., 2023; Zhang et al., 2024): our characterization runs the classical pathwise–probabilistic translation in the necessity direction. *Prediction sets to decisions* (Prinster et al., 2026; Wang & Dobriban, 2026; Hu et al., 2026): the decision layer composes the static certified-risk interfaces with the SIP ledgers. *Deployment controllers* (three anonymized manuscripts): each is a projection of the SIP framework (Section 3.4).

2 The Sequential Intervention Process

We model deployment as a sequential process: at each step the controller chooses an intervention, the delivered output is booked once against a running budget, and feedback matures with delay. The protocol fixes the objects and the intervention dichotomy; the axioms then discipline the ledger and the gate on structural change.

2.1 Protocol

We consider a stream $y_1, y_2, \ldots$ arriving on the filtration $(\mathcal{F}_t)_{t\geq 0}$. At each step t a deployed pipeline in state Π_t can be run under any intervention a from a finite library $\mathcal{A}$, producing an output $O_t(a)$ (a prediction set, or an abstention token). The controller observes evidence $e_t \in \mathcal{F}_{t-1}$, selects $a_t \in \mathcal{A}$ at cost $c(a_t) \geq 0$, and the user receives $O_t(a_t)$. Feedback matures with delay at most H: the loss $\ell_t = L(O_t(a_t), y_t) \in [0, B]$ is observable by $t + H$. We then charge a risk proxy $\tilde{\ell}_t \in [0, 1]$ of the *delivered* output (e.g. the miscoverage indicator, or a bounded monotone risk) to a budget with target level $\alpha \in (0, 1)$.

Definition 2.1 (Intervention dichotomy)**.** The library splits as $\mathcal{A} = \mathcal{A}_{\mathrm{C}} \cup \mathcal{A}_{\mathrm{S}}$. A *calibration actuator* $a \in \mathcal{A}_{\mathrm{C}}$ modifies only the served threshold $q_t \in \mathbb{R}$ (widening, quantile tracking, level shifts); it does not change the pipeline state. A *structural actuator* $a \in \mathcal{A}_{\mathrm{S}}$ changes the pipeline state Π_t (input repair, model switch, fallback to a baseline, abstention, operator scheduling); it may change the conditional law of all future losses arbitrarily.

The dichotomy is load-bearing: calibration actuators admit trajectory-wise guarantees through a scalar recursion whose telescoping does not care how the scores were generated; structural actuators change the sequence itself and require stochastic certification.

2.2 Axioms

Three axioms discipline the loop: a recursion on the served threshold (A1), single booking of every step against the delivered output (A2), and an e-process gate on structural state changes (A3).

Axiom 2.2 (Budget recursion)**.** The served threshold evolves *unclipped* on $\mathbb{R}$, $q_{t+1} = q_t + \gamma(\tilde{\ell}_t - \alpha) + u_t$, where $\gamma > 0$, $\tilde{\ell}_t \in [0,1]$ is the risk proxy of the delivered output, and the drift u_t is $\mathcal{F}_{t-1}$-measurable with $|u_t| \le u_{\max}$ per step and $|\sum_{t \le T} u_t| \le U$ for all T. With maturity delay $H > 1$ the recursion consumes proxies in maturity order—the update at step t applies to the oldest unbooked matured proxy $\tilde{\ell}_s$, $s \le t - H$ — so every step is still booked exactly once (A2). The telescoping identity is then unchanged up to the at most H not-yet-matured steps, which add $\le H$ to the numerator slack. (Equivalently, index the recursion by maturity time; the e-processes of A3 use only matured statistics by the same convention.) When the delivered output is an abstention token, its proxy is fixed at $\tilde{\ell}_t = \alpha$ (budget-neutral): abstaining neither earns nor spends risk budget, so the ledger cannot be gamed by abstaining forever.

We keep the recursion unclipped: clipping would break the telescoping identity on which the deterministic ledger rests. The unclipped state is the bookkeeping variable; the served output maps it onto the library's feasible range (an interval half-width $\max\{\epsilon_q, q_t\}$, the nearest feasible quantile level), so that clipping acts on the served side and never touches the state. Uniform-in-T control instead requires a feasibility property of the *library*:

Assumption 2.3 (Boundary feasibility)**.** There exist finite $\underline{q} \le \bar{q}$ such that on every trajectory $\tilde{\ell}_t \le \alpha$ whenever $q_t \ge \bar{q}$ and $\tilde{\ell}_t \ge \alpha$ whenever $q_t \le \underline{q}$: past either boundary the next increment of A1 cannot push further out. (For interval widening the stronger form $\tilde{\ell}_t = 0$ above $\bar{q}$ and $\tilde{\ell}_t = 1$ below $\underline{q}$ holds whenever the library contains a "predict everything" output: e.g. bounded scores, $\bar{q} =$ score bound. The inequalities stated are all the proofs use, and they are the form compatible with budget-neutral abstentions, whose proxy is pinned at α by A1; for abstention and fallback outputs the condition must still be checked.)

Axiom 2.4 (Single booking)**.** Each step is booked exactly once, against the output the user received: the proxy $\tilde{\ell}_t$ entering A1 is computed from $O_t(a_t)$ (post-action), and no step is booked twice or retroactively re-booked after a later intervention. The intervention a_t is $\mathcal{F}_{t-1}$-measurable.

Axiom 2.5 (E-gated structural change)**.** Each structural state change $a \in \mathcal{A}_{\mathrm{S}}$ carries a null hypothesis H_0^a: $\mathbb{E}[D_t^a \mid \mathcal{F}_{t-1}] \le 0$ whenever H_0^a holds at t, where $D_t^a \in [-B_D, B_D]$ is an observable paired shadow statistic (e.g. the loss difference between the current pipeline and the post-change pipeline evaluated on the same matured outcome). Each gate updates only on matured statistics (D_s^a with $s \le t - H$). The change may fire at t only if a nonnegative-supermartingale-under-H_0^a e-process E_t^a exceeds $1/\alpha_j$, where the spending schedule satisfies $\sum \alpha_j \le \alpha_{\mathrm{S}}$ jointly across all segments of *all* gates (e.g. each of G gates receives budget α_{S}/G with per-segment spending $\alpha_j \propto 1/j(j+1)$), and each e-process restarts on a window schedule of length W.

A1 generalizes the ACI recursion, the conformal PID loop, and the online CRC recursion; A2 axiomatizes post-action accounting (the ledger tracks what was delivered, not what the un-intervened pipeline would have delivered); A3 generalizes the single-null e-gated fallback to a family of nulls over the structural library. One scope point: everything the deterministic ledger controls is the *booked proxy* $\tilde{\ell}_t$—an observable $[0,1]$ score of the delivered output chosen by the operator—so if the proxy is a poor surrogate for the harm the operator cares about, the ledger faithfully controls the wrong thing.

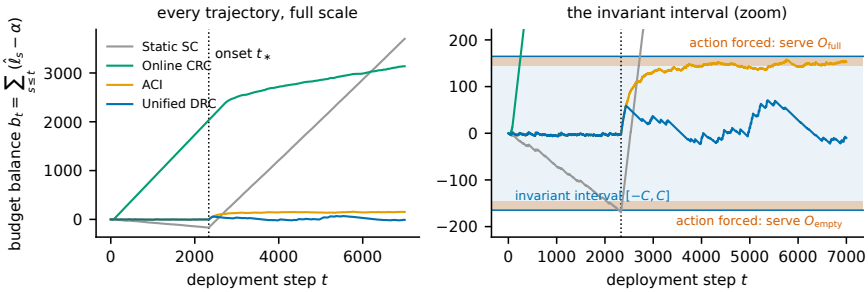

Figure 2: The invariant scalar on a real stream (E2: NP, degraded, seed 0; protocol in Section 5). *Left:* the budget balance b_t for four policies; the static threshold (after onset t_*) and online-CRC (controlling the graded proxy it recurses on, not the booked indicator) leave every bounded interval. *Right, zoom:* the A1-recursion policies stay confined to the invariant interval $[-C, C]$ of Theorem A(I), implied by the ledger identity $b_T = (q_{T+1} - q_1 - \sum_t u_t)/\gamma$ (Theorem 3.1(a)). In the boundary strips (schematic width), part (II) forces every compliant policy's output; ACI rides near the upper boundary, while the full library holds the balance well inside.

3 The Characterization, and the Theorem that Achieves It

We first answer the structural question (*what do all compliant controllers look like?*), and then construct one: the characterization delimits the design space, while the two-ledger theorem realizes the invariant for the full library.

3.1 The Design Space: A Characterization

We first strip the SIP to its accounting core: scores $s_t \in [0, 1]$ arrive adversarially, the served output either covers s_t or not, $\tilde{\ell}_t = \mathbf{1}\{s_t \text{ not covered}\}$, and the *budget balance* $b_T := \sum_{t \le T}(\tilde{\ell}_t - \alpha)$, $b_0 = 0$, records booked proxy against budget; under A1, $b_T = (q_{T+1} - q_1 - \sum_{t \le T} u_t)/\gamma$, so the two-ledger theorem will confine this scalar. A policy is *trajectory-compliant at level* C, $P(C)$, if $|b_T| \le C$ for every score sequence and T; O_{full} covers every score ($\tilde{\ell}_t = 0$ surely), O_{empty} covers none ($\tilde{\ell}_t = 1$ surely).

Theorem A (Invariant characterization)**.** *Let π be a deterministic policy over a library containing O_{full} and O_{empty}.*

(I) Budget invariant. *The following are equivalent:* (i) *π satisfies $P(C)$ for some finite C;* (ii) *for some finite C', π serves O_{full} whenever $b_{t-1} > C' - (1 - \alpha)$ and O_{empty} whenever $b_{t-1} < -C' + \alpha$;* (iii) *for some finite C'' and every $\eta > 0$, both $\exp(\eta(b_t - C''))$ and $\exp(-\eta(b_t + C''))$ are e-processes for the universal null (the family of* all *distributions over score sequences). (Explicit constant translations between C, C', C'' are proved in Appendix B.)*

(II) Boundary necessity. *If π satisfies $P(C)$, then after every reachable history:* (a) *if $b_{t-1} > C - (1 - \alpha)$, π must serve O_{full};* (b) *if $b_{t-1} < -C + \alpha$, π must serve an output covering no score;* (c) *if the library lacks O_{full} or lacks O_{empty}, no deterministic policy satisfies $P(C)$ for any finite C.*

In words, part (I) says that budget compliance *is* confinement of a single scalar to an invariant interval; part (II) says that at its edges compliance removes all freedom, and a library missing an extreme output is uncontrollable at any finite radius.

Proof sketch. Necessity (II) is an adversary argument: if above the upper strip the policy served any output leaving some score s^* uncovered, the adversary replays the reachable history and then plays s^*, forcing $b_t = b_{t-1} + (1 - \alpha) > C$; the lower strip is symmetric, and a library missing an extreme output loses a fixed increment at every step, giving (c). Sufficiency ((ii)⇒(i)) is induction: the forced boundary outputs produce increments of the restoring sign ($-\alpha$ above, $1 - \alpha$ below) while any interior step moves the balance by less than 1, so b_t never leaves $[-C' - 1, C' + 1]$. The equivalence (i)⇔(iii) is a pathwise translation: $|b_t| \le C$ iff both exponential processes are bounded by one on every path, and a nonnegative adapted process is an e-process for the universal null exactly when it is bounded by one pathwise. Full proofs, constant translations, and extensions are in Appendix B. Part

(II) is why the two-ledger design below is canonical: at the boundaries every compliant controller must act identically, so controllers differ only strictly inside—where the deterministic ledger moves the balance and the intervention ledger prices the structural actuators' discrete jumps. The universal-null e-process fact underlying (i)⇔(iii) (Ruf et al., 2023; Zhang et al., 2024) and the graded-proxy extension are discussed there as well. Figure 2 shows both regimes on a real stream: non-compliant balances leave every bounded interval after degradation onset while the A1 policies stay confined.

3.2 THE TWO-LEDGER THEOREM

Two ledgers running side by side suffice to hold the invariant for the full library.

Theorem 3.1 (Deployment risk control)**.** *Fix any intervention policy π, arbitrarily history-dependent (including adversarially chosen triggers). Then:*

(a) Deterministic risk ledger. *Under Axioms 2.2 and 2.4, on every trajectory and for every $T \geq 1$,*

$$\Big|\frac{1}{T}\sum_{t=1}^{T}\tilde{\ell}_t - \alpha\Big| \;\leq\; \frac{|q_{T+1} - q_1| + U}{\gamma\, T},$$

regardless of which structural interventions fired and when. Under Assumption 2.3 in addition, $q_t \in [\underline{q} - \gamma - u_{\max} - 2U,\ \bar{q} + \gamma + u_{\max} + 2U]$ for all t, so the bound is uniform:

$$\Big|\frac{1}{T}\sum_{t=1}^{T}\tilde{\ell}_t - \alpha\Big| \;\leq\; \frac{\bar{q} - \underline{q} + 2(\gamma + u_{\max}) + 5U}{\gamma\, T}.$$

If the budget is maintained per stratum g over an $\mathcal{F}_{t-1}$-measurable partition, the same bounds hold within each stratum with T replaced by the stratum count T_g.

(b) Anytime-valid intervention ledger. *Under Axioms 2.2, 2.4 and 2.5, with gates updated on matured statistics only (each gate at time t uses D_s^a with $s \leq t - H$),*

$$\mathbb{P}\big(\exists\, a \in \mathcal{A}_{\mathrm{S}},\ \exists \text{ segment } j :\ H_0^a \text{ holds throughout segment } j \text{ and gate } a \text{ fires in segment } j\big) \;\leq\; \alpha_{\mathrm{S}},$$

where the spending $\sum \alpha_j \leq \alpha_{\mathrm{S}}$ runs over all segments of all gates. The bound is trajectory-level and hence valid at every data-dependent stopping time simultaneously, with no continuous-monitoring correction. Note, however, that it controls firings in segments where the null holds throughout*; it does not—and cannot—constrain a firing at an instant where the null momentarily holds after being violated earlier in the same segment.*

Remark 3.2 (Detection power). The intervention ledger (b) is complemented by a power statement: under a detectability condition (†) on the window length, a warranted change fires within $W + O(W)$ steps of onset with probability $\geq 1 - \beta$; below the Δ-vs-W frontier the windowed gate may never fire (the detectability-frontier theorem in Appendix C).

The proof (Appendix C) assembles three techniques. For (a), the unclipped recursion telescopes exactly, $q_{T+1} - q_1 = \gamma \sum_{t \leq T}(\tilde{\ell}_t - \alpha) + \sum_{t \leq T} u_t$: an algebraic identity of the delivered trajectory, not a probabilistic statement. A *single-booking lemma* (from A2) makes the identity survive structural interventions: each step contributes exactly one proxy term computed from the delivered output, so a switch or repair changes future proxy *values* but never the step-to-entry correspondence. A confinement lemma under Assumption 2.3 bounds $|q_{T+1} - q_1|$: excursions beyond $[\underline{q}, \bar{q}]$ book increments pointing back toward the interval. For (b), on each restart segment the wealth is a nonnegative supermartingale under that segment's null, so Ville's inequality (Ville, 1939) bounds the false-firing probability by the segment's spend α_j, and $\sum \alpha_j \leq \alpha_{\mathrm{S}}$ gives the uniform statement. Part (a) is to be read at its stated scope: an identity about the booked proxy of the delivered trajectory, not the deployment's unobserved true risk.

3.3 THE LEDGERS ARE COUPLED, NOT PARALLEL

The two ledgers of Theorem 3.1 are not independent: every structural change shocks the calibration recursion. Under per-action booking, the served threshold decomposes as $q_t = \hat{q}(\pi_t) + o_t$, so

each pipeline switch jumps the base quantile by at most $\kappa = \max_{\pi,\pi' \in \Pi_{\text{reach}}} |\hat{q}(\pi) - \hat{q}(\pi')|$ (a deployment-time measurable constant), while the structure-free drift budget is U_0. We call an instant *degraded* if the fallback gate's null fails there and *strictly benign* if the recovery gate's null fails; we let M_T count the maximal degraded episodes started by T and N_T the structural changes.

Theorem 3.3 (Coupled ledgers)**.** *Run the two-gate fallback/recovery protocol—the fallback and recovery gate pair every gated experiment runs, formalized in Appendix C.7—under Axioms 2.2, 2.4 and 2.5 and Assumption 2.3, with per-action booking and matured feedback. Then with probability at least* $1 - \alpha_{\text{S}}$*,* simultaneously for all $T \geq 1$*:* (i) *every fallback firing by* T *occurs in a gate segment containing a degraded instant, and every recovery firing in a segment containing a strictly benign instant;* (ii) $N_T \leq 2M_T$*;* (iii) *the deterministic ledger holds with an environment-determined (not intervention-determined) constant:*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \tilde{\ell}_t - \alpha \right| \;\leq\; \frac{\bar{q} - \underline{q} + 2(\gamma + u_{\max} + \kappa) + 5\,(U_0 + \kappa)}{\gamma\, T}.$$

The two ledgers certify different things. The deterministic ledger's constant is self-contained: cumulative structural drift telescopes to the quantile diameter κ, so the bound survives arbitrary switching—even $N_T = \Theta(T)$ — with no gate at all. But no deterministic ledger can see whether a firing was warranted or what it cost (Theorem 3.4): ungated, a trigger can make $\Theta(T)$ false interventions at positive cost each, with no certificate. The intervention ledger closes this hole: on the Ville good event, firings occur only in segments containing the right kind of instant, and $N_T \leq 2M_T$ bounds activity by the environment's episode structure, not the policy's. A cost extension (structural cost bounded by episode durations plus $4W$ per episode) is stated and proved in Appendix C.7. The episode-counting argument is specific to the two-gate protocol; the general-library version is open.

3.4 Special cases recovered

By restricting the triple $(\mathcal{A}, \tilde{\ell}, \alpha_{\text{S}})$ we recover—with the same constants—the guarantees of ACI, the bounded-drift subfamily of conformal PID, and online CRC (R1–R3; formal statements in Appendix D). R4–R6 cover the three anonymized deployment controllers of the related-work taxonomy (stratified debt control, operator orchestration, e-gated fallback); since those manuscripts cannot be independently verified in print, the recoveries live only in Appendix D. Together with the decision-layer recoveries (Appendix F), the externally checkable recoveries carry the subsumption claim on their own.

3.5 Failure boundaries

The axioms are necessary, not decorative; the representative boundary is the gate's:

Theorem 3.4 (No free lunch for ungated structural change)**.** *Fix* $\alpha_{\text{S}} = \infty$ *(no gate). Then no trigger rule admits a nontrivial uniform bound on its false-intervention frequency and its missed-detection delay simultaneously: for every trigger rule and every* $\epsilon > 0$ *there is an environment on which either the false-intervention frequency stays bounded away from zero (with cumulative cost growing linearly), or a warranted intervention is missed for an arbitrarily long stretch. Throughout, the deterministic ledger still holds on every trajectory; it simply certifies nothing about intervention validity.*

Two further boundaries delimit the remaining axioms (statements and proofs in Appendix E): booking a step's proxy against the *pre*-intervention output while the user receives the post-intervention one (violating A2 with A1 intact) lets the realized delivered risk stay a constant above budget forever; and an e-process gate without restarts must first repay its accumulated benign-period log-wealth debt—detection delay $\Omega(\rho \lambda \mu\, t^*)$ after onset, capped at $O(W)$ by A3's windowed restarts.

In words: an ungated trigger cannot buy validity and power together; A2's post-action accounting and A3's restarts are likewise load-bearing, each removed by an explicit counterexample.

4 THE DECISION LAYER

The ledgers just established certify properties of the served *sets*; deployments are judged by the decisions taken on top of them. The certificates carry downstream without new distributional assumptions.

4.1 DECISION INTERFACE AND THE BRIDGE INEQUALITY

The interface prices each served interval by its certified best action.

Definition 4.1 (Decision interface)**.** We define a *decision interface* on top of the SIP as a triple $(\mathbb{A}, \mathcal{D}, \ell^{\mathrm{dec}})$: a nonempty action space $\mathbb{A}$; a measurable map $\mathcal{D}$ taking the served prediction set C_t (the interval $[f_t - q_t, f_t + q_t]$, or the abstention token) to an action $z_t = \mathcal{D}(C_t) \in \mathbb{A}$; and a decision loss $\ell^{\mathrm{dec}} : \mathbb{A} \times \mathcal{Y} \to [0, B_{\mathrm{dec}}]$, $B_{\mathrm{dec}} < \infty$. Following Wang & Dobriban (2026), the *certified robust risk* of action z against set C at level α is

$$L_C(z; \alpha) \;=\; \ell_C(z) + \alpha\big(\bar{\ell}(z) - \ell_C(z)\big)_+,$$

where $\ell_C(z) = \sup_{y \in C} \ell^{\mathrm{dec}}(z, y)$ and $\bar{\ell}(z) = \sup_y \ell^{\mathrm{dec}}(z, y)$. We write $R_t = L_{C_t}(\mathcal{D}(C_t); \alpha)$ for the served certificate; R_t is *observable at prediction time* (it depends only on the served set and the loss, not on y_t).

Theorem 4.2 (Decision bridge and ledger)**.** (i) Bridge inequality. *For every step t, on every trajectory, and for any history-dependent intervention policy,*

$$\ell^{\mathrm{dec}}(z_t, y_t) \;\le\; R_t + B_{\mathrm{dec}}\, \mathbf{1}\{y_t \notin C_t\}.$$

(ii) Decision ledger. *Instantiate Axioms 2.2 and 2.4 with $\tilde{\ell}_t = \mathbf{1}\{y_t \notin C_t\}$. Then on every trajectory and for every $T \ge 1$,*

$$\frac{1}{T}\sum_{t=1}^{T} \ell^{\mathrm{dec}}(z_t, y_t) \;\le\; \frac{1}{T}\sum_{t=1}^{T} R_t + B_{\mathrm{dec}}\big[\alpha + \tfrac{|q_{T+1} - q_1| + U}{\gamma T}\big],$$

and under Assumption 2.3 the bracket is $\alpha + O(1/\gamma T)$ uniformly. Under the coupled protocol, with probability at least $1 - \alpha_{\mathrm{S}}$ simultaneously for all T, the bracket's numerator becomes the constant of Theorem 3.3(iii).

The bridge (i) is pointwise and distribution-free (proof in Appendix F): if the set covers, the realized loss is at most the certified worst case; if it misses, the indicator pays. Composing pathwise with Theorem 3.1(a) gives (ii). The decomposition is the operational content: average *certified* risk is the observable price of the interface, $B_{\mathrm{dec}}\,\alpha$ the irreducible out-of-set allowance, the residual slack the deployment-time constant $U_0 + \kappa$ under the coupled protocol. The decision layer thereby inherits the intervention ledger's certification with no exchangeability, stationarity, or knowledge of the data law. On scope: we presume outputs are sets, so budget-neutral abstentions sit outside the ledger.

Two corollaries are deferred to Appendix F: Corollary F.2 bounds the interventional decision regret of the minimax map $\mathcal{D}_\alpha$, and Corollary F.3 shows the newsvendor minimax certificate is linear in served width, converting the compliance-at-width frontier into a decision-loss frontier. The layer extends further (Appendix G). When the deployment must also choose *which* of K base policies to serve, a meta-policy tracks the best fixed library member to within $O\big(\sqrt{(H{+}1)\ln(K/\delta)}/\sqrt{T}\big)$ anytime; a gated variant makes every adoption carry the same e-process certificate as a base-level structural intervention; and one composite ledger bounds all materialized interventions, base or meta, at $\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}}$ uniformly over stopping times. All remaining statements, discussion, and proofs are in Appendices F and G.

5 EXPERIMENTS

We compute all numbers here and in the appendix from the released raw outputs using the released analysis scripts; no number appears only in the paper. "Holds the budget"/"compliance" means the empirical booked proxy (coverage of the delivered intervals), not any unobserved true risk.

Table 1: Pooled over six streams (mean over per-stream seed-means ± sd across streams; 10 seeds each). *cov.* = empirical coverage (budget $1 - \alpha = 0.90$); *width* = mean served width (calibration-sd units); *cost* = total intervention cost; *shortfall* = mean/max over the 60 runs of $(0.90 - \text{coverage})_+$ (threshold sensitivity in Table 2).

	clean				degraded			
policy	cov.	width	cost	shortfall	cov.	width	cost	shortfall
Static SC	0.915 ±0.038	0.28	0	0.010/0.055	0.459 ±0.144	0.28	0	0.441/0.535
ACI (R1)	0.900 ±0.001	0.27	0	0.000/0.001	0.884 ±0.007	1.65	0	0.016/0.022
DtACI	0.899 ±0.002	0.28	0	0.002/0.005	0.895 ±0.002	1.70	0	0.005/0.011
Betting ACI	0.887 ±0.001	0.25	0	0.013/0.016	0.879 ±0.003	1.66	0	0.021/0.024
Conformal PI (R2)	0.900 ±0.000	0.28	0	0.000/0.000	0.897 ±0.001	1.76	0	0.003/0.005
Online CRC (R3)	0.029 ±0.034	0.01	0	0.871/0.891	0.421 ±0.079	0.99	0	0.479/0.658
e-Fallback (R6)	0.915 ±0.038	0.28	0	0.010/0.055	0.872 ±0.065	0.53	744	0.041/0.129
Debt-Repair (R4)	0.900 ±0.001	0.27	44	0.000/0.001	0.884 ±0.007	1.64	67	0.016/0.022
Unified DRC	0.900 ±0.001	0.27	44	0.000/0.001	0.900 ±0.003	0.65	770	0.001/0.006

Setup. We use six real hourly streams: five day-ahead electricity-price markets (NP, PJM, BE, FR, DE) from the open EPF benchmark (Lago et al., 2021) and the ETTh1 transformer-temperature stream (Zhou et al., 2021), under four backbone families: a linear AR model (main results), a two-layer MLP, and two zero-shot foundation models (Chronos-Bolt, TimesFM-2.5). The structural library contains input repair (rolling median, applied in *every* scenario, so clean-stream repairs are real operations) and fallback to a seasonal-naive baseline; calibration is split-conformal. Scenarios: *clean*; *degraded* (after onset t^* the model's input view is corrupted with 1.5σ noise, the baseline's view stays clean); *degraded-both* (both views corrupted, removing fallback's best-case advantage). Feedback delay $H = 1$; ten seeds per configuration; full protocol in Appendix I. We index the experiments E1–E15, each a protocol fixed before execution; per-experiment tables, breakdowns, and tests are in Appendix H (numbering follows the protocols: E8b inside E8; E9 and E11 inside E7).

Policies under one budget. All policies share $(\alpha, \alpha_S) = (0.1, 0.05)$, the same calibration data, and the same matured-feedback protocol: `noop` (static split conformal), the recovered special cases `aci` (R1), `conformal_pi` (R2's bounded-drift member), `crc_online` (R3), `efallback` (R6), `debt_repair` (R4), and `unified` (full library under A1–A3). We add two calibration-only baselines, `aci_dt` (DtACI (Gibbs & Candès, 2024)) and `aci_bet` (ACI-by-betting (Podkopaev et al., 2024)), the two strongest published upgrades of the fixed-step recursion ported to our radius representation (raw rows ship in the revision package); neither can repair, fall back, or act structurally.

Budget compliance and its price (E1–E2). Under injected degradation, compliance alone does not separate the field: the unified policy, conformal PI, and DtACI all sit at max shortfall ≤ 0.011 over 60 runs (Table 1). What separates them is the price of compliance: the unified policy holds the budget 2.5–2.7× narrower than every coverage-targeting alternative, on all six streams (Holm-adjusted cluster-bootstrap $p < 10^{-3}$, stream level; exact Wilcoxon $p = 0.031$ at the n=6 floor), with lower graded risk. Figure 2 already showed the mechanism on a real degraded stream, and Figure 5 follows the served intervals step by step. On clean streams the unified policy pays a repair overhead with zero false gate firings; online CRC controls exactly the graded functional it recurses on while its coverage collapses. The frontier replicates under a shallow-MLP backbone (every failure mode replicates) and, correctly, *vanishes* under two zero-shot foundation-model backbones—the injected noise is absorbed by the backbone itself, and the gates stay silent until a real feed failure arrives.

Where the library helps, and where it is dead weight (E1, E3, E4). Corrupt *both* input views and fallback's advantage vanishes: the unified policy's structural spend becomes pure overhead. Natural symmetric regime breaks are the opposite extreme: no e-gate fires ($M_T = 0$; Table 6). The two recorded upstream feed failures fire the fallback gate on both (Table 7). The unified policy holds the budget at $5.5\times$ narrower width on the electricity outage. On the more severe traffic outage, where no policy holds 0.90 over the full deployment, it attains near-best event coverage at $1.6\times$ narrower.

The decision layer (E5, E5c). Routing every policy's intervals through the newsvendor and storage interfaces converts the width frontier into a decision frontier (Figure 3; full values in Table 14), as the width-linearity of the minimax certificate predicts. Under degradation the unified policy attains the lowest realized newsvendor loss at a certificate-violation rate of 0.100—the budget α itself. The day-level block ledger (a joint ledger over D-step violation events) pulls every policy to the block budget.

The meta and composite ledgers (E6–E9, E11). The meta layer's premise is real: the best library member changes with the scenario in every backbone. Within it, Hedge tracks below the anytime bound in all 360 runs, GCL (the gate-confirmed leader; Appendix G) adopts zero switches on clean

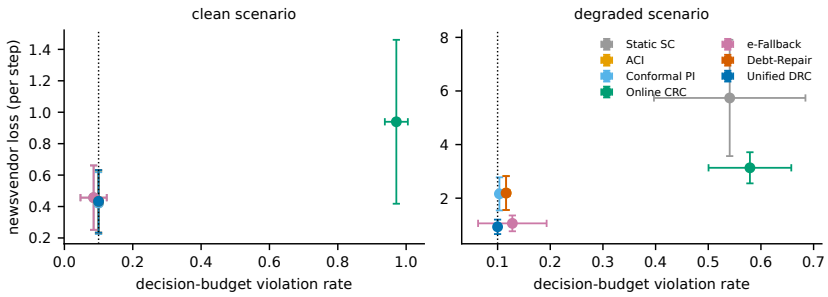

Figure 3: **The decision frontier.** Realized newsvendor loss versus decision-certificate violation rate (E5; pooled over applicable streams and 10 seeds; bars show sd across streams). The dotted line marks the decision budget $\alpha = 0.10$. Clean streams are decision-indistinguishable; under degradation, unified DRC is the only policy in the lower-loss, on-budget corner.

streams, and no clean run materializes an intervention (a served action) under either composite-budget convention (shared α_S, or rebudgeted α_S/K).

Two real external deployments, replayed (E10, E12). The CDC FluSight hub (2,470 cells; Figure 8) and the US COVID-19 Forecast Hub (2,610 cells) are replayed retrospectively. The gates fire only on models genuinely behind the hub baseline (one power miss; size zero on both hubs), and from each hub's own ensemble the certified answer is restraint—zero GCL adoptions on both hubs, with its multi-scale variant MS-GCL certifying a single near-tied switch on FluSight.

Independent external baselines (E13). E13 rules out a built-in width advantage head-to-head against three methods that are *not* special cases on the identical E1 protocol (Table 27): SPCI (Xu & Xie, 2023), NexCP (Barber et al., 2023), and PH-recal (Page–Hinkley (Page, 1954; Gama et al., 2014)); all recalibration-layer methods, none able to repair inputs or fall back. On clean streams the externals sit within 13% of ACI or undercover. Under degradation only NexCP holds the budget, at 2.4–2.7× the unified width (significant, stream level, both backbones); SPCI and PH-recal breach it. With both views corrupted the advantage collapses to $\leq 15\%$; NexCP holds the budget and the unified policy does not.

A decision task with externally defined loss (E14). E5's loss parameters are ours, so E14 borrows a task verbatim from the external conformal robust optimization (CRO) literature (Hu et al., 2026; Sun et al., 2023): single-period storage arbitrage on the five EPF streams (minimax map, certificate, and robustness functional; protocol hash-anchored before execution). At the same $\alpha = 0.10$ the R9 ledger (the robustness-control objective; Appendix F) pins the realized violation frequency at 0.0997–0.1002 on all five streams while coverage-route policies over-satisfy it (0.016–0.045), converting the slack into roughly twice the realized arbitrage profit. One registered sub-hypothesis failed: static CRO drifted conservative rather than violating.

Beyond forecasting (E15). The deterministic ledger is not tied to interval forecasting: on UCI Covertype's spatially drifting row order, a frozen classifier serves score-threshold label sets; over 521,012 steps the static threshold drifts to 3.7× the budget while the A1 recursion holds the booked proxy at 0.1011; one registered guardrail failed.

6 Limitations and Conclusion

Limitations (full list in Appendix J). (i) guarantees concern the booked proxy $\tilde{\ell}_t$ and the gates' nulls under boundary feasibility (Assumption 2.3), not unobserved harm; (ii) the detectability frontier assumes a persistent effect size; (iii) only two structural actuators are exercised; ungated repairs lie outside the certificates; (iv) foundation-model evidence is two models; (v) meta tracking holds on detectable classes and assumes no performativity (Assumption G.1, relaxed quantitatively by Proposition G.10); (vi) E10/E12 are retrospective replays. **Conclusion.** Deployment-time intervention on served prediction systems is invariance control of the running risk-budget balance, with no choice at the boundaries. One theorem gives both ledgers, recovering ACI, conformal PID, and online CRC as projections; under injected degradation on real streams the full library is alone on the compliance-at-width frontier, and on two external deployments the certified answer was restraint.

Ethics statement. This paper advances the reliability of deployed forecasting systems; better-audited intervention accounting benefits safety-critical downstream users, though the guarantees concern booked risk proxies, not real-world harm; no negative societal consequence is foreseen.

Reproducibility statement. All experiments are CPU-level and deterministic given the released seeds; the released material ships source, raw results, and scripts regenerating all tables and figures; proofs are in Appendices B–G, the artifact layout in Appendix I, and an anonymized mirror accompanies the submission.

AI use statement. Generative AI assisted drafting and polishing of the manuscript and development of the experiment harness; all output was reviewed and verified against independent computation. AI was not used for research ideas, data, or results; the authors take responsibility for all content.

REFERENCES

Anastasios N. Angelopoulos, Emmanuel Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Anonymous. Coverage debt: post-action accounting for stratified conformal deployment. *Under review*, 2027a. Anonymized third-party reference; details withheld for double-blind review.

Anonymous. Anytime-valid fallback for deployed forecasting models. *Under review*, 2027b. Anonymized third-party reference; details withheld for double-blind review.

Anonymous. No-regret orchestration of repair and calibration operators for deployed forecasters. *Under review*, 2027c. Anonymized third-party reference; details withheld for double-blind review.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

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.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, 2023.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

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

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

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

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yang Hu, Jieren Tan, Changliang Zou, Yajie Bao, and Haojie Ren. Conformal robustness control: A new strategy for robust decision. In *International Conference on Learning Representations*, 2026. OpenReview forum `bt4Ahpemmi`.

Varun Kotte. When can conformal risk control certify LLM outputs? bounds, impossibility, and adaptation for structured generation. arXiv:2606.29054, 2026.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Tuo Liu, Edgar Dobriban, and Francesco Orabona. Online conformal prediction via universal portfolio algorithms. arXiv:2602.03168, 2026.

Jeffery Opoku and David Banahene. Promptshift-crc: Drift-aware conformal risk control for foundation models under prompt and domain shift. arXiv:2606.15964, 2026.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2022.

Aleksandr Podkopaev, Dong Xu, and Kuang-Chih Lee. Adaptive conformal inference by betting. In *International Conference on Machine Learning*, pp. 40886–40907, 2024.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. *arXiv preprint arXiv:2603.02196*, 2026.

Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1704–1722, 2017.

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

Johannes Ruf, Martin Larsson, Wouter M. Koolen, and Aaditya Ramdas. A composite generalization of Ville's martingale theorem using e-processes. *Electronic Journal of Probability*, 28:1–21, 2023.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A*, 184(2):407–431, 2021.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, 2019.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2(2): 229–260, 2024.

Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual lp. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sophia Huiwen Sun and Rose Yu. Conformal prediction for time-series forecasting with change points. In *Advances in Neural Information Processing Systems*, 2025.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: conformal test martingales for change-point detection. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, pp. 191–210, 2021.

Tao Wang and Edgar Dobriban. Optimal decision-making based on prediction sets. *arXiv preprint arXiv:2602.00989*, 2026.

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

Marcelo J. Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, 2023.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866, 2022.

Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang. On the existence of powerful p-values and e-values for composite hypotheses. *Annals of Statistics*, 52(5), 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

A Extended Related Work

Online conformal recalibration. ACI (Gibbs & Candès, 2021) guarantees $|T^{-1}\sum_t \mathrm{err}_t - \alpha| = O(1/\gamma T)$ for arbitrary sequences via a one-dimensional recursion; extensions tune the step size, aggregate experts, or set it by betting (Zaffran et al., 2022; Bhatnagar et al., 2023; Gibbs & Candès, 2024; Podkopaev et al., 2024) (the expert-aggregation and betting variants are benchmarked in E1; Appendix H.1), and conformal PID control (Angelopoulos et al., 2023) adds integrator and scorecaster modules. We keep γ fixed by design: the ledger's target is the *unweighted* booked frequency, for which a constant step size is what makes the telescoping identity hold unchanged—the $O(1/\gamma T)$ constant is the price of that invariance, while adaptive-step variants control regret-type or weighted functionals, a different ledger. Our deterministic ledger (Theorem 3.1a) is the common generalization: any bounded-drift recursion on the *delivered* threshold inherits the bound under single booking (A2), no matter what structural interventions permute the score sequence. Change-point-aware conformal methods (Sun & Yu, 2025) take a different route to non-stationarity: a learned switching dynamical system predicts state transitions and consolidates per-state forecasts into the delivered interval, anticipating shifts rather than reacting to them. In the SIP dichotomy this is still a calibration actuator (the delivered set changes, the pipeline state does not), so it could serve inside $\mathcal{A}_{\mathrm{C}}$ under the same ledger; the e-gates of Axiom 2.5 answer a different question, whether a structural intervention is warranted at all, and the two compose rather than compete.

Risk control beyond coverage. CRC (Angelopoulos et al., 2024) replaces the coverage indicator by any bounded monotone risk under exchangeability; Feldman et al. (2023) give the online analogue. Drift-aware variants reweight the calibration pool and adjust the level online (Opoku & Banahene, 2026): adaptive calibration layers in our terminology, with no budget ledger and no converse characterization of the policies that hold one. Kotte (2026) bound when a *static* CRC certificate is feasible at all for structured generation; their impossibility results concern one-shot certification, not the sequential intervention loop whose failure boundaries we delimit in Section 3.5. Theorem 3.1(a) holds for any $[0,1]$-bounded risk proxy of the delivered output and Proposition D.3 recovers the online CRC guarantee; we do not claim the exchangeability-based batch theorem, whose assumptions the deployment setting drops by construction (Remark C.4).

Anytime-valid sequential testing. E-processes and test martingales (Ville, 1939; Shafer, 2021; Grünwald et al., 2024; Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024) give type-I control uniform over data-dependent stopping, used for risk monitoring (Podkopaev & Ramdas, 2022), change detection (Shin et al., 2024; Vovk et al., 2021), and model fallback (Anonymous, 2027b). Axiom A3 lifts the single-null construction to a *family* of nulls over the structural library with a shared spending schedule; Theorem 3.1(b) is the resulting uniform guarantee.

Sufficiency versus necessity. Pathwise-to-probabilistic translations are classical: regret–martingale equivalences (Rakhlin & Sridharan, 2017) and defensive forecasting (Vovk et al., 2005; Shafer & Vovk, 2019). Our characterization (Section 3.1) is of this family but runs in the *necessity* direction for a control problem: it delimits what every compliant controller must do. The universal-null e-process fact it uses is essentially known (Ruf et al., 2023; Zhang et al., 2024); we cite it rather than claim it. Recent online conformal advances (Liu et al., 2026) strengthen sufficiency (parameter-free coverage); boundary necessity bounds the complementary side.

Prediction sets to decisions. Conformal policy control (Prinster et al., 2026), optimal decision-making with prediction sets (Wang & Dobriban, 2026), and conformal robustness control (Hu et al., 2026) connect set-valued predictions to downstream decision quality: all assuming a *static* pipeline. Our decision layer (Section 4) adopts the certified-robust-risk interface of Wang & Dobriban (2026) and shows it composes with the SIP ledgers: decision-loss control survives history-dependent interventions, and the static objectives are recovered when the library is trivial. The delimitation from Hu et al. (2026) is exact: CRC optimizes the *construction* of a static prediction set under an explicit robustness constraint on i.i.d. data (batch training plus i.i.d.-assumed test-time calibration); the SIP holds the set constructor fixed and controls the *sequential intervention trajectory* of a deployed pipeline; their constraint functional, transplanted to the ledger, is R9 (Proposition F.6).

Deployment controllers. Three recent controllers each solve one slice: stratified debt accounting (Anonymous, 2027a), no-regret repair orchestration with anytime e-certificates (Anonymous, 2027c), and e-gated fallback with windowed restarts (Anonymous, 2027b). Section 3.4 shows each is a

projection of the SIP framework; Section 3.5 shows the boundaries they hit (ungated repairs, wealth debt) are fundamental, not incidental.

B THE CHARACTERIZATION: DISCUSSION AND PROOFS

The proofs are collected below, with randomized versions (Theorems B.2 and B.3: almost-sure or probability-$(1-\varepsilon)$ compliance forces boundary compliance up to the same exceptional budget) and a counterexample showing *expectation*-level compliance forces nothing: the invariant is a trajectory-level object. The equivalence (i)⇔(iii) rests on a known fact: a universal-null e-process is exactly a process bounded by one on every path (Ruf et al., 2023; Zhang et al., 2024). Part (c) is the binary shadow of Assumption 2.3: absent the extreme outputs, no policy is compliant. (Binary miscoverage proxy; graded proxies spanning $\{0,1\}$ force the same extremes with scaled thresholds.)

B.1 SETTING

We work in the accounting core of Section 3.1. Scores $s_1, s_2, \cdots \in [0,1]$ are chosen by an adversary. At each step t the policy serves an output from a library $\mathcal{O}$; each output $O \in \mathcal{O}$ determines a coverage set $\mathrm{cov}(O) \subseteq [0,1]$, the booked proxy is $\tilde{\ell}_t = \mathbf{1}\{s_t \notin \mathrm{cov}(O_t)\}$, and the budget balance is $b_T = \sum_{t \le T}(\tilde{\ell}_t - \alpha)$ with $b_0 = 0$. O_{full} denotes an output with $\mathrm{cov} = [0,1]$ and O_{empty} one with $\mathrm{cov} = \emptyset$. A *deterministic* policy π maps each finite score prefix to an output; a *randomized* policy additionally uses internal randomness, and O_t is measurable with respect to the past scores $s_{1:t-1}$ and the policy's coins; in particular O_t does not see s_t. For a deterministic policy, a history is *reachable* if some score prefix produces it; since the adversary may play any prefix, every reachable history is available to the lower-bound constructions below. Property $P(C)$: $|b_T| \le C$ for every score sequence and every $T \ge 1$. Throughout, increments satisfy $\tilde{\ell}_t - \alpha \in \{-\alpha,\, 1-\alpha\}$.

B.2 A KNOWN LEMMA: UNIVERSAL-NULL E-PROCESSES ARE PATHWISE BOUNDS

The following equivalence is a special (and easy) case of the composite generalization of Ville's theorem; see Ruf et al. (2023) for the general theory of e-processes for composite families and Zhang et al. (2024) for existence characterizations. We state and prove it only to keep the appendix self-contained; no novelty is claimed.

Lemma B.1 (Universal null; known)**.** *Let $(E_t)_{t\ge 0}$ be a nonnegative process adapted to the score filtration (no auxiliary randomization), and let $\mathcal{P}_{\mathrm{all}}$ be the family of all probability distributions over score sequences, including all point masses. Then (E_t) is an e-process for $\mathcal{P}_{\mathrm{all}}$ (i.e. $\mathbb{E}_P[E_\tau] \le 1$ for every $P \in \mathcal{P}_{\mathrm{all}}$ and every stopping time τ) if and only if $\sup_{\omega,t} E_t(\omega) \le 1$.*

Proof. If $E_t \le 1$ pathwise, then $\mathbb{E}_P[E_\tau] \le 1$ for every P and τ trivially. Conversely, fix any path ω and any t; take $P = \delta_\omega$ (the point mass on ω) and $\tau \equiv t$. Since E_t is a function of the scores alone, $\mathbb{E}_P[E_\tau] = E_t(\omega) \le 1$. (Theorem A concerns deterministic policies, for which b_t (hence each process in (iii)) is adapted to the score filtration, so the lemma applies as stated.) □

B.3 PROOF OF THEOREM A(II) (BOUNDARY NECESSITY, DETERMINISTIC)

Proof. (a) Suppose, for contradiction, that after some reachable history h with balance $b_{t-1} > C-(1-\alpha)$ the policy serves an output O_t with $\mathrm{cov}(O_t) \ne [0,1]$, i.e. there exists $s^* \in [0,1]\setminus \mathrm{cov}(O_t)$. Because π is deterministic, the adversary can present the score prefix realizing h (reachability) followed by $s_t = s^*$. Then $\tilde{\ell}_t = 1$ and

$$b_t \;=\; b_{t-1} + (1-\alpha) \;>\; C,$$

violating $P(C)$. Hence $\mathrm{cov}(O_t) = [0,1]$: the policy must serve a full-prediction output.

(b) Symmetrically, suppose after a reachable history with $b_{t-1} < -C + \alpha$ the served output covers some $s^* \in \mathrm{cov}(O_t)$. The adversary plays the realizing prefix followed by $s_t = s^*$; then $\tilde{\ell}_t = 0$ and

$$b_t \;=\; b_{t-1} - \alpha \;<\; -C,$$

violating $P(C)$. Hence $\mathrm{cov}(O_t) = \emptyset$.

(c) Suppose the library contains no output with $\mathrm{cov} = [0,1]$. Then at every step, whatever π serves leaves some score uncovered; the adversary plays such a score at every step, so $\tilde{\ell}_t = 1$ for all t and $b_T = (1-\alpha)T \to \infty$, defeating every finite C. If the library contains no output with $\mathrm{cov} = \emptyset$, the adversary plays a covered score at every step, so $\tilde{\ell}_t = 0$ for all t and $b_T = -\alpha T \to -\infty$. In either case no deterministic policy satisfies $P(C)$ for any finite C. □

B.4 Proof of Theorem A (budget invariant)

Proof. (i) ⇒ (ii) with $C' = C$. Immediate from Theorem A(II)(a)–(b); an output covering no score behaves identically to O_{empty} for the ledger, so we may take it to be O_{empty}. (If $C < \frac{1}{2}$ the two boundary zones overlap; Theorem A(II) then shows no balance in the overlap is reachable under a compliant policy, so the forcing conditions remain consistent.)

(ii) with $C' \geq \frac{1}{2}$ ⇒ (i) at level $C'+1$. The restriction $C' \geq \frac{1}{2}$ ensures the two boundary zones $\{b > C' - (1-\alpha)\}$ and $\{b < -C' + \alpha\}$ are disjoint, so the hypothesis of (ii) is consistent. We show $b_t \in [-C'-1,\, C'+1]$ for all t by induction; the base case is $b_0 = 0$. Assume $b_{t-1} \in [-C'-1, C'+1]$ and consider three exhaustive cases. If $b_{t-1} > C' - (1-\alpha)$, the policy serves O_{full}, so $\tilde{\ell}_t = 0$ and $b_t = b_{t-1} - \alpha$, giving $b_t \leq C' + 1 - \alpha < C' + 1$ and $b_t > C' - 1 \geq -C' - 1$. If $b_{t-1} < -C' + \alpha$, the policy serves O_{empty}, so $\tilde{\ell}_t = 1$ and $b_t = b_{t-1} + (1-\alpha)$, giving $b_t \geq -C' - 1 + (1-\alpha) > -C' - 1$ and $b_t < -C' + 1 \leq C' + 1$. Otherwise $b_{t-1} \in [-C' + \alpha,\, C' - (1-\alpha)]$ and the increment lies in $(-1, 1)$, so $b_t \in (-C' + \alpha - 1,\, C' - (1-\alpha) + 1) \subseteq (-C' - 1,\, C' + 1)$. Hence π satisfies $P(C'+1)$.

(i) ⇔ (iii) with $C'' = C$. If $|b_t| \leq C$ on every path, then $\exp(\eta(b_t - C)) \leq 1$ and $\exp(-\eta(b_t + C)) \leq 1$ on every path for every $\eta > 0$; both processes are nonnegative and adapted, hence e-processes for $\mathcal{P}_{\mathrm{all}}$ by Lemma B.1. Conversely, if both are e-processes for $\mathcal{P}_{\mathrm{all}}$; Lemma B.1 gives $\exp(\eta(b_t - C'')) \leq 1$ and $\exp(-\eta(b_t + C'')) \leq 1$ pathwise, i.e. $b_t \leq C''$ and $b_t \geq -C''$ for every path and t; hence $P(C'')$. □

B.5 Randomized policies

For randomized policies we specialize, as in the online conformal literature, to the threshold library: outputs are $\{[0,w] : w \in [0,1]\} \cup \{O_{\mathrm{empty}}\}$, coverage of s means $s \leq w$, O_{full} is $w = 1$, and w_t is measurable with respect to $s_{1:t-1}$ and the policy's coins. The adversary is *oblivious* (fixes the score sequence in advance) unless stated otherwise.

Theorem B.2 (B-R1: almost-sure ledger forces almost-sure boundary compliance)**.** *Let π be a randomized policy such that for every score sequence, $\mathbb{P}(\forall T :\ |b_T| \leq C) = 1$. Then for every score sequence and every t:* (a) $\mathbb{P}\big(b_{t-1} > C - (1-\alpha)$ *and* $w_t < 1\big) = 0$*;* (b) $\mathbb{P}\big(b_{t-1} < -C + \alpha$ *and* $O_t \neq O_{\mathrm{empty}}\big) = 0$.

Proof. (a) Fix a score sequence σ and a step t, and suppose for contradiction that $p := \mathbb{P}(b_{t-1} > C - (1-\alpha),\ w_t < 1) > 0$ under σ. Since $\{w_t < 1\} = \bigcup_n \{w_t \leq 1 - 1/n\}$, there is $\delta > 0$ with $p' := \mathbb{P}(b_{t-1} > C - (1-\alpha),\ w_t \leq 1 - \delta) > 0$. Both b_{t-1} and w_t are functions of $s_{1:t-1}$ and the coins only, so this probability is unchanged under the modified sequence σ' that agrees with σ up to $t-1$ and sets $s_t = 1 - \delta/2$. Under σ', on that event $s_t > w_t$, so $\tilde{\ell}_t = 1$ and $b_t = b_{t-1} + (1-\alpha) > C$: the ledger is violated with probability $\geq p' > 0$ under σ', contradicting almost-sure compliance on every sequence.

(b) Any output other than O_{empty} is a threshold set $[0, w_t]$ with $w_t \geq 0$ and therefore covers $s = 0$. If $p' := \mathbb{P}(b_{t-1} < -C + \alpha,\ O_t \neq O_{\mathrm{empty}}) > 0$ under σ, modify σ to set $s_t = 0$: on that event $\tilde{\ell}_t = 0$ and $b_t = b_{t-1} - \alpha < -C$ with probability $\geq p' > 0$, the same contradiction. □

Theorem B.3 (B-R2: high-probability ledger forces high-probability boundary compliance)**.** *Let $\varepsilon \in [0,1)$ and let π be a randomized policy such that $\mathbb{P}(\forall T :\ |b_T| \leq C) \geq 1 - \varepsilon$ against every* adaptive *score adversary (one that chooses s_t as a function of the realized outputs and proxies before t). Fix any $\delta > 0$ and call step t a* boundary deviation *if either $b_{t-1} > C - (1-\alpha)$ and $w_t \leq 1 - \delta$, or $b_{t-1} < -C + \alpha$ and $O_t \neq O_{\mathrm{empty}}$. Then there is an adaptive adversary under which*

$$\mathbb{P}\big(\exists\, t :\ \textit{step } t \textit{ is a boundary deviation}\big) \ \leq\ \varepsilon.$$

Under this adversary (the one that punishes every deviation) boundary compliance at margin δ fails on an event of probability at most ε: the same exceptional budget the ledger itself is allowed.

Proof. Consider the adversary that plays $s_t = 1 - \delta/2$ whenever $b_{t-1} > C - (1-\alpha)$, plays $s_t = 0$ whenever $b_{t-1} < -C + \alpha$, and plays $s_t = 0$ otherwise (any fixed choice works there). This adversary is adaptive: b_{t-1} is a function of the realized proxies, which it observes. Let τ be the first boundary-deviation step (possibly ∞). On $\{\tau < \infty\}$: if the deviation is at the upper boundary, $w_\tau \le 1 - \delta < s_\tau$, so $\tilde{\ell}_\tau = 1$ and $b_\tau = b_{\tau-1} + (1-\alpha) > C$; if at the lower boundary, $s_\tau = 0 \le w_\tau$, so $\tilde{\ell}_\tau = 0$ and $b_\tau = b_{\tau-1} - \alpha < -C$. Either way $\{\tau < \infty\} \subseteq \{\exists T : |b_T| > C\}$, whose probability is at most ε by hypothesis. □

Remark B.4 (Expectation-level compliance forces nothing). The trajectory-level (or probability-level) hypothesis in Theorems B.2 and B.3 cannot be weakened to expectations. Consider the policy that flips a single coin at $t = 0$: with probability α it serves O_{empty} forever, with probability $1 - \alpha$ it serves O_{full} forever. For every score sequence, $\tilde{\ell}_t \equiv 1$ on the first branch and $\tilde{\ell}_t \equiv 0$ on the second, so

$$\mathbb{E}[b_T] \;=\; \alpha(1-\alpha)T + (1-\alpha)(-\alpha)T \;=\; 0 \qquad \text{for all } T,$$

the strongest possible expectation-level ledger. Yet on the first branch (probability $\alpha > 0$) the balance $b_t = (1-\alpha)t$ crosses every boundary and the policy serves a non-full output at every step above it: the expected number of boundary deviations is infinite. In particular no finite deviation budget follows from $\sup_T |\mathbb{E}[b_T]| \le C$; Theorem B.3 is the correct probability-level substitute. The budget invariant is a statement about trajectories, not about means.

B.6 Positioning

The deterministic-to-stochastic translation in Theorem A(i)⇔(iii) belongs to a known family: Rakhlin & Sridharan (2017) prove an equivalence between pathwise regret inequalities and martingale tail bounds, and defensive forecasting (Vovk et al., 2005; Shafer & Vovk, 2019) converts game-theoretic pathwise statements into probabilistic ones. Our use is the elementary corner of that family (the universal null; Lemma B.1), and the contribution of this appendix is not the translation but the *necessity* direction of Theorem A(II) and Theorems B.2 and B.3: sufficiency results in online conformal prediction (from ACI (Gibbs & Candès, 2021) to universal-portfolio methods (Liu et al., 2026)) construct policies whose budget balance is confined; the theorems here show every confined policy, randomized or not, is forced to the extreme outputs at the boundary, and that the boundary-feasibility assumption (Assumption 2.3) is therefore necessary for every policy class. The counterexample of Remark B.4 separates trajectory-level from expectation-level accounting; the latter is genuinely weaker and supports no boundary characterization.

C The Two-Ledger Theorem and the Coupled-Ledger Extension

Theorem 3.1's two ledgers are not independent: every structural change shocks the calibration recursion; the served base threshold jumps to the new pipeline's calibration quantile, a structure following from Axiom 2.4 alone. Lemma C.5 below bounds the drift by $u_{\max} + \kappa$ per step and $U_0 + \kappa$ cumulatively (κ a quantile diameter measurable at deployment time; the per-switch jumps telescope, so the cumulative bound is independent of the switch count N_T); Definition C.6 defines the environment's degraded-episode count M_T. Theorem 3.3 (stated in Section 3.3; cost extension Theorem C.8 stated and proved in Appendix C.7) proves, for the two-gate fallback/recovery protocol of Proposition D.6: with probability $\ge 1 - \alpha_S$, simultaneously for all T, (i) every fallback (recovery) firing occurs in a segment containing a degraded (benign) instant; (ii) $N_T \le 2M_T$; (iii) the deterministic ledger of Theorem 3.1(a) holds with the constant $U_0 + \kappa$, independent of both the policy's activity and M_T; (iv) under per-episode detectability (†), cumulative structural cost is bounded by the episode durations plus $4W$ per episode. The coupling runs both ways but asymmetrically: the deterministic ledger needs no gate (its constant is immune to switching by telescoping), yet it certifies nothing about intervention validity—with $\alpha_S = \infty$ an ungated trigger can make $\Theta(T)$ false interventions with unbounded cumulative cost and no validity certificate (Theorem 3.4); and without Assumption 2.3, the pipeline need not keep serving long enough for the gates' matured statistics to accumulate. What the intervention ledger certifies is what the deterministic ledger cannot: firing validity and the episode-controlled cost.

C.1 Remarks on the two-ledger theorem

Theorem C.1 (Excess-risk and cost ledger under a detected episode (power complement to Theorem 3.1))**.** *In the setting of Theorem 3.1, suppose a degradation of persistent effect size $\Delta > 0$ (i.e. $\mathbb{E}[D_t^a \mid \mathcal{F}_{t-1}] \geq \Delta$ after onset t^*) begins after a benign period, and that the window is long enough to detect it:*

$$W \;\geq\; C\,\frac{B_D^2}{\Delta^2}\Big(\log\frac{K}{\alpha_{\mathrm{S}}\,\beta} + 2\log\big(\tfrac{t^*}{W}+2\big)\Big), \tag{$\dagger$}$$

for an absolute constant C. Then with the mixture-betting gate of Axiom 2.5 over a K-point grid, the warranted change fires within $W + O(W)$ steps of onset with probability $\geq 1-\beta$, and the cumulative excess risk before the change is at most the per-step excess times that delay; total intervention cost is bounded by $\sum_t c(a_t)$, which the policy observes and can budget online. If ($\dagger$) fails, the windowed gate may never *fire: each restart resets the wealth to 1 while the threshold $1/\alpha_j$ grows, so degradations below the Δ-vs-W detectability frontier are permanently invisible (Remark C.3).*

Remark C.2 (Boundary feasibility is a real failure boundary). Without Assumption 2.3 no uniform bound is possible: a structural degradation may make the pipeline miscover at every finite threshold ($\tilde{\ell}_t = 1$ for all q), in which case q_t grows linearly and the trajectory bound $|q_{T+1} - q_1|/\gamma T = \Theta(1)$ is tight; the ledger residual then reports that the budget is not being met. A *clipped* recursion in the same environment pins at its ceiling and its telescoping bound is simply false. Boundary feasibility is thus itself a condition that structural actuators can violate, extending the failure-boundary catalogue of Section 3.5.

Remark C.3 (Detectability frontier). Condition ($\dagger$) is not an artifact: windowed restarts trade the wealth debt of Theorem E.2 for blindness to effects below $\Delta \approx B_D\sqrt{C\log(K/\alpha_{\mathrm{S}}\beta)/W}$. Small persistent degradations under this frontier are permanently invisible to the gated detector, and only the deterministic ledger (a) (which keeps charging their excess risk) registers them. The spending schedule also bounds the detection *horizon*: at per-episode confidence $\beta_i \propto 1/i^2$ (Theorem C.8) the required window count n of Theorem C.1 grows like $2\log i$ in the episode label, so for fixed W and margin Δ there is a largest certifiable episode index beyond which ($\dagger$) cannot be met—the power guarantee is effectively a finite-horizon statement, buying per-episode power early at the price of late episodes.

Remark C.4 (What is not claimed). Theorem 3.1(a) is a long-run frequency guarantee on the realized trajectory, the same type as ACI's; it is not a finite-sample distributional guarantee. We do not recover the batch CRC theorem of Angelopoulos et al. (2024), whose $\mathbb{E}[\ell] \leq \alpha$ holds under exchangeability: an assumption deployment-time distribution shift violates by its nature; Proposition D.3 recovers its *online* counterpart (Feldman et al., 2023). Distribution-free finite-sample *conditional* (per-instance) risk control is impossible (Foygel Barber et al., 2021), and the framework does not claim it.

C.2 The coupled-ledger shock structure

The shock structure is not an extra assumption: it is a consequence of Axiom 2.4's per-action booking.

Lemma C.5 (Structural shock from per-action booking)**.** *Under Axiom 2.4 with per-action calibration over a finite structural library; the served threshold decomposes as $q_t = \hat{q}(\pi_t) + o_t$, where $\hat{q}(a)$ is action a's calibration quantile (frozen at deployment) and the offset o_t follows the Axiom 2.2 recursion with structure-free drift budget $|\sum_{t\leq T}\tilde{u}_t| \leq U_0$ and $|\tilde{u}_t| \leq u_{\max}$; the combined drift u_t of the served threshold satisfies, writing N_T for the number of structural state changes up to time T,*

$$|u_t| \leq u_{\max} + \kappa \quad \textit{and} \quad \Big|\sum_{t\leq T} u_t\Big| \leq U_0 + \kappa \;\; \textit{for all } T, \tag{1}$$

with $\kappa = \max_{\pi,\pi'\in\Pi_{\mathrm{reach}}}|\hat{q}(\pi) - \hat{q}(\pi')|$, where Π_{reach} is the set of reachable pipeline states *(the default (model) pipeline together with every pipeline reachable through the structural library $\mathcal{A}_{\mathrm{S}}$) so that κ covers all jumps the served state π_t can make, not only those between actuators. It is a deployment-time measurable constant (all per-action calibration quantiles $\hat{q}(\pi)$, $\pi \in \Pi_{\mathrm{reach}}$, are frozen at deployment). Proof in Appendix C.7.*

Definition C.6 (Episodic environment)**.** Call t a *degraded* instant if the fallback null fails at t ($\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] > 0$) and a *strictly benign* instant if the recovery null fails at t ($\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] < 0$). Let M_T be the number of maximal degraded episodes (maximal runs of degraded instants separated by strictly benign time) started by time T, and d_i the length of episode i.

Remark C.7 (Sharpness of the coupling). The episode dependence in (ii) and the cost bound is not an artifact: a degraded episode can force one fallback firing and one recovery firing, so $N_T \le 2M_T$ is tight up to constants, and the per-episode $4W$ overhead of Theorem C.8 is the window-restart price. The risk-ledger constant carries no episode dependence at all: a fallback–recover round trip shifts the base quantile by at most κ each way, but the two shifts cancel in the cumulative drift $\hat{q}(\pi_{T+1}) - \hat{q}(\pi_1)$, so M_T round trips still cost only κ in the risk numerator. The unified policy of Section 5 instantiates the theorem with κ = the measured gap between per-action calibration quantiles.

The coupled-ledger theorem itself (Theorem 3.3) is stated in Section 3.3; its cost extension is:

Theorem C.8 (Coupled ledgers: cost extension)**.** *In the setting of Theorem 3.3, if in addition, for every episode i,* both *the episode's degradation effect* and *the ensuing benign stretch's recovery effect satisfy the detectability condition* (†) *at the per-episode confidence $\beta_i = (\beta/2)/i(i+1)$ (so $\sum_i 2\beta_i \le \beta$, a spending schedule mirroring Axiom 2.5), then with probability at least $1 - \alpha_{\mathrm{S}} - \beta$ (a fixed constant, not involving the random M_T) the cumulative structural cost obeys*

$$\sum_{t \le T} c(a_t) \;\le\; \sum_{i \le M_T} c_{\mathrm{sw}}\big(d_i + 4W\big),$$

where c_{sw} is the per-step cost of serving the fallback pipeline (Lemma C.10 makes the interval accounting explicit).

Two points sharpen the asymmetry stated above. First, no redundancy in accounting substitutes for the gate: two parallel deterministic ledgers would still certify nothing about firing validity, and the ungated trigger of Theorem 3.4 exploits this. Second, the cost bound couples in the other direction: the deterministic ledger's feasibility (Assumption 2.3) keeps the pipeline serving, rather than abstaining, long enough for the gates' matured statistics to accumulate. Proof in Appendix C.7.

C.3 Single-booking lemma

Lemma C.9 (Single booking)**.** *Under Axiom 2.4, the sequence $(\tilde{\ell}_t)_{t \ge 1}$ entering the recursion of Axiom 2.2 equals the sequence of risk proxies of the outputs delivered to the user, one term per step, and this identification is unaffected by which interventions fired at any step.*

Proof. By Axiom 2.4, at each step t exactly one output $O_t(a_t)$ is delivered, exactly one proxy $\tilde{\ell}_t = \tilde{\ell}(O_t(a_t), y_t)$ is computed from it, and exactly this proxy enters the recursion at step t; no step is booked twice and no booked step is later modified. Hence the map $t \mapsto \tilde{\ell}_t$ is a bijection between steps and booked entries, and each entry is the delivered proxy. Structural interventions change *which* output is delivered at future steps and therefore change the *values* $\tilde{\ell}_s$ for $s > t$, but not the identification between booked entries and delivered proxies. □

The lemma is where naive composition fails: if a later repair re-books an earlier step (or a step is booked against the un-repaired output while the repaired one is delivered), the recursion and the delivered risk decouple, and Theorem E.1 shows the gap is $\Theta(1)$, not vanishing.

C.4 Proof of Theorem 3.1(a)

Proof. Trajectory bound. The recursion of Axiom 2.2 is unclipped, so it telescopes exactly:

$$q_{T+1} - q_1 = \gamma \sum_{t=1}^{T} (\tilde{\ell}_t - \alpha) + \sum_{t=1}^{T} u_t.$$

By Lemma C.9 the $\tilde{\ell}_t$ are the delivered proxies regardless of the structural intervention history, so the identity holds on every trajectory. Rearranging and using $|\sum_t u_t| \le U$,

$$\Big|\sum_{t=1}^{T} (\tilde{\ell}_t - \alpha)\Big| \le \frac{|q_{T+1} - q_1| + U}{\gamma},$$

using $|\sum_t u_t| \le U$. Dividing by T gives the first display of Theorem 3.1(a).

Confinement under Assumption 2.3. We show $q_t \le \bar{q} + \gamma + u_{\max} + 2U$ for all t (the lower bound $q_t \ge \underline{q} - \gamma - u_{\max} - 2U$ is symmetric); w.l.o.g. $q_1 \in [\underline{q}, \bar{q}]$. Consider any excursion above $\bar{q}$: let σ be the last step with $q_\sigma \le \bar{q}$, so $q_{\sigma+1} \le \bar{q} + \gamma(1-\alpha) + u_{\max} \le \bar{q} + \gamma + u_{\max}$. For every subsequent step s of the excursion, $q_s > \bar{q}$, hence $\tilde{\ell}_s \le \alpha$ by Assumption 2.3 and the increment is $q_{s+1} - q_s = \gamma(\tilde{\ell}_s - \alpha) + u_s \le u_s$. Summing over the excursion of length m,

$$q_{\sigma+1+m} - q_{\sigma+1} \;=\; \gamma \sum_{s=\sigma+1}^{\sigma+m} (\tilde{\ell}_s - \alpha) + \sum_{s=\sigma+1}^{\sigma+m} u_s \;\le\; \sum_{s=\sigma+1}^{\sigma+m} u_s \;\le\; 2U,$$

since every window sum of the drift is bounded by $2U$ (difference of two partial sums, each $\le U$ in absolute value). Hence $q_t \le \bar{q} + \gamma + u_{\max} + 2U$ throughout the excursion, and the claim follows. Therefore $|q_{T+1} - q_1| \le \bar{q} - \underline{q} + 2(\gamma + u_{\max} + 2U)$; substituting into the trajectory bound gives numerator $\bar{q} - \underline{q} + 2(\gamma + u_{\max}) + 5U$, the uniform bound of Theorem 3.1(a) (for the drift-free special cases $u \equiv 0$, used by R1/R3, the constant reduces to $(\bar{q} - \underline{q} + 2\gamma)/\gamma$, matching the ACI constant).

Strata. If a separate recursion state q_t^g is maintained for each stratum g of an $\mathcal{F}_{t-1}$-measurable partition, and booking is per stratum (each step booked once, in its own stratum, against the delivered output), then the argument above applies within each stratum with horizon $T_g = |\{t \le T : g_t = g\}|$. □

C.5 Proof of Theorem 3.1(b)

Proof. Fix a gate $a \in \mathcal{A}_{\mathrm{S}}$ and a restart segment j of its e-process, with assigned budget α_j. Define the event N_j^a = "H_0^a holds throughout segment j". On N_j^a, the process $E_t^{a,j} = \prod_s (1 + \lambda_s D_s^a)$ (mixture over a finite grid of $\lambda \in [0, 1/(2B_D)]$, so each factor is positive; increments use matured statistics D_s^a, $s \le t - H$, which are $\mathcal{F}_{t-1}$-measurable) is a nonnegative supermartingale: $\mathbb{E}[1 + \lambda_s D_s^a \mid \mathcal{F}_{s-1}] \le 1$ whenever $\mathbb{E}[D_s^a \mid \mathcal{F}_{s-1}] \le 0$, and a mixture of supermartingales is a supermartingale. By Ville's inequality (Ville, 1939),

$$\mathbb{P}\big(N_j^a \text{ and } \sup_t E_t^{a,j} \ge 1/\alpha_j\big) \le \alpha_j.$$

A structural change fires only upon such a crossing (Axiom 2.5). Note the null is well-posed within each segment because the served pipeline is frozen off the firing times: a structural action changes the served pipeline only at a crossing, which by construction ends the current segment and restarts the gate, so throughout any single segment j the pipeline whose degradation H_0^a speaks about is fixed, and the supermartingale above is constructed under that fixed pipeline; the coupling between firing and pipeline change never invalidates a segment's null mid-segment. A union bound over all segments of all gates, with $\sum \alpha_j \le \alpha_{\mathrm{S}}$ jointly, gives

$$\mathbb{P}\big(\exists\, a, j : N_j^a \text{ and gate } a \text{ fires in segment } j\big) \le \alpha_{\mathrm{S}}.$$

This is exactly the event in the statement of Theorem 3.1(b). Note the guarantee is per-segment-null: if the null is violated early in a segment (wealth grows legitimately) and later holds again, a firing at such an instant is not (and cannot be) controlled; the statement excludes it by design. The bound is a probability statement about the entire trajectory, hence valid at every data-dependent stopping time simultaneously; no continuous-monitoring correction is needed. □

C.6 Proof of Theorem C.1 (part (c) of the two-ledger guarantee)

Proof. After onset, $\mathbb{E}[D_t^a \mid \mathcal{F}_{t-1}] \ge \Delta$. Consider the first full window beginning after onset; the wait for it is at most W. Within that window the e-process restarts with unit wealth. For the grid point λ^* closest to the Kelly-optimal $\Delta/(2B_D^2)$ (within a factor of 2 by the geometric grid), each factor satisfies $\mathbb{E}[\log(1 + \lambda^* D_t^a) \mid \mathcal{F}_{t-1}] \ge c_0 \Delta^2 / B_D^2$ for an absolute constant c_0 (second-order expansion of $\log(1+x)$ with $|\lambda^* D| \le 1/2$). The mixture wealth dominates $1/K$ times the λ^*-wealth. Azuma–Hoeffding on the bounded log-increments gives that after

$$n = O\Big(\frac{B_D^2}{\Delta^2}\Big(\log\frac{K}{\alpha_{\mathrm{S}}\beta} + 2\log\big(\tfrac{t^*}{W} + 2\big)\Big)\Big),$$

post-onset steps within the window, $\log E_t \ge \log(1/\alpha_j)$ with probability $\ge 1 - \beta$. Explicitly: the segment containing onset is $j \le t^*/W + 1$ (one segment per window of W steps over a horizon t^*),

and with $\alpha_j \propto \alpha_S/j(j+1)$,

$$\log \frac{1}{\alpha_j} \;=\; \log \frac{j(j+1)}{\alpha_S} \;\le\; \log \frac{1}{\alpha_S} + 2\log(j{+}1) \;\le\; \log \frac{1}{\alpha_S} + 2\log\left(\tfrac{t^*}{W} + 2\right),$$

which together with the $\log(K/\beta)$ mixture-and-confidence term is the n displayed. The detectability condition (†) guarantees $n \le W$, so the crossing occurs within the first full post-onset window; without (†) the argument fails irreparably: each restart resets the wealth to 1 while the threshold $1/\alpha_j$ keeps growing, so a gate whose per-window wealth growth never reaches the threshold never fires, for any number of windows (Remark C.3). Total delay $\le W + n \le 2W$. The cumulative pre-change excess risk is (per-step excess, bounded by $\bar{\delta} \le 1$) times the delay, and the cost ledger $\sum_t c(a_t)$ is observable by construction. □

C.7 Proof of Theorem 3.3 (coupled ledgers, with the cost extension Theorem C.8)

Proof of Lemma C.5. Write the served-threshold update between consecutive steps: $q_{t+1} - q_t = (o_{t+1} - o_t) + (\hat{q}(\pi_{t+1}) - \hat{q}(\pi_t))$. The offset follows the Axiom 2.2 recursion, $o_{t+1} - o_t = \gamma(\tilde{\ell}_t - \alpha) + \tilde{u}_t$, so the served threshold follows the same recursion with combined drift $u_t = \tilde{u}_t + (\hat{q}(\pi_{t+1}) - \hat{q}(\pi_t))$. The jump term is nonzero only at structural state changes; the state π_t ranges over Π_{reach} (the default pipeline and the states reachable through $\mathcal{A}_S$), so each jump is bounded by $\kappa = \max_{\pi,\pi' \in \Pi_{\text{reach}}} |\hat{q}(\pi) - \hat{q}(\pi')|$ in absolute value, and $|u_t| \le u_{\max} + \kappa$. Cumulatively the jumps telescope exactly: $\sum_{t \le T} \big(\hat{q}(\pi_{t+1}) - \hat{q}(\pi_t)\big) = \hat{q}(\pi_{T+1}) - \hat{q}(\pi_1)$, bounded by κ regardless of the switch count N_T — round trips cancel rather than accrue. Hence $|\sum_{t \le T} u_t| \le |\sum_{t \le T} \tilde{u}_t| + \kappa \le U_0 + \kappa$. Both $\hat{q}$ and Π_{reach} are fixed at deployment, so κ is measurable from the calibration data alone. □

Proof of Theorem 3.3. Let G be the complement of the event of Theorem 3.1(b): on G, no gate fires in a segment where its null holds throughout, and $\mathbb{P}(G) \ge 1 - \alpha_S$. G is an event about the whole trajectory, so everything derived from it below holds simultaneously for all T.

(i). The fallback gate's null is H_0^{fb}: $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] \le 0$; it holds throughout a segment iff the segment contains no degraded instant (Definition C.6). On G, therefore, every fallback firing's segment contains a degraded instant. Symmetrically, the recovery gate's null H_0^{rec}: $\mathbb{E}[-D_t \mid \mathcal{F}_{t-1}] \le 0$ holds throughout a segment iff the segment contains no strictly benign instant, so on G every recovery firing's segment contains a strictly benign instant.

(ii). The protocol is a two-state machine starting in the model state: firings strictly alternate fallback, recovery, fallback, …, so $N_T \le 2F_T$ where F_T is the number of fallback firings by T. We claim that on G, distinct fallback firings correspond to distinct degraded episodes, whence $F_T \le M_T$ and $N_T \le 2M_T$. Let $F^{(1)} < F^{(2)}$ be consecutive fallback firing times. By alternation there is a recovery firing time R with $F^{(1)} < R < F^{(2)}$. By (i), $F^{(1)}$'s segment contains a degraded instant $t_1 \le F^{(1)}$; R's segment starts after $F^{(1)}$ (the recovery gate is (re)started when the fallback fires) and ends at R, so it contains a strictly benign instant t_b with $F^{(1)} < t_b \le R$; and $F^{(2)}$'s segment starts after R, so it contains a degraded instant t_2 with $R < t_2$. Thus $t_1 < t_b < t_2$ with t_b strictly benign: t_1 and t_2 lie in different maximal degraded episodes. Distinct consecutive fallback firings are therefore charged to distinct episodes, proving the claim.

(iii). Repeat the proof of Theorem 3.1(a) with the drift budget Equation (1) supplied by Lemma C.5. The telescoping identity is unchanged; the trajectory bound uses $|\sum_{t \le T} u_t| \le U_0 + \kappa$ (the per-switch jumps telescope), and the confinement argument uses window sums $|\sum_{s=\sigma+1}^{\sigma+m} u_s| \le 2(U_0 + \kappa)$ (difference of two partial sums, each bounded by $U_0 + \kappa$) and per-step drift $u_{\max} + \kappa$. The numerator of the uniform bound is therefore $\bar{q} - \underline{q} + 2(\gamma + u_{\max} + \kappa) + 5(U_0 + \kappa)$, with no dependence on N_T or M_T: the risk-ledger constant is immune to switching because the jumps cancel in the telescoping sum. This holds pathwise on every trajectory, for every T at once; (iii) needs neither (i) nor (ii)—which is the asymmetry discussed after Theorem 3.3: the gate's job is validity and cost, not the risk constant.

Cost bound. Enumerate episodes in order of onset ($i = 1, 2, \ldots$; a deterministic indexing of a random sequence). For episode i, let A_i be the event that the fallback gate, if in the model state at onset, fires within $2W$ steps of onset s_i^{ep} (if the system is instead already serving the fallback at onset, episode i requires no firing: by the interval-accounting invariant of Lemma C.10, the active serving interval is

charged to the earlier episode that triggered it, so no interval is double-counted and no episode is left uncharged; A_i then holds vacuously), and B_i the event that the recovery gate, once monitoring during the post-episode benign stretch, fires within $2W$ steps of the later of (episode end e_i^{ep}, its own segment start). By the two-sided detectability hypothesis; Theorem C.1 applied to each gate gives $\mathbb{P}(A_i^c) \le \beta_i$ and $\mathbb{P}(B_i^c) \le \beta_i$ with $\beta_i = (\beta/2)/i(i+1)$. The events are indexed by the deterministic label i, not by the random count M_T, so the countable union bound is legitimate and

$$\mathbb{P}\Big(\bigcup_{i\ge 1}(A_i^c \cup B_i^c)\Big) \;\le\; \sum_{i\ge 1} 2\beta_i \;=\; \beta \sum_{i\ge 1} \tfrac{1}{i(i+1)} \;=\; \beta,$$

a fixed constant. On $G \cap \bigcap_i (A_i \cap B_i)$; Lemma C.10 below charges each maximal fallback-serving interval to a distinct episode with length at most $d_i + 4W$; the per-step cost of serving the fallback is c_{sw} and no other structural cost is charged by this protocol, giving $\sum_{t\le T} c(a_t) \le \sum_{i \le M_T} c_{\mathrm{sw}}(d_i + 4W)$ with probability $\ge 1 - \alpha_{\mathrm{S}} - \beta$. □

Lemma C.10 (Interval accounting). *On $G \cap \bigcap_i (A_i \cap B_i)$, each maximal fallback-serving interval $[s, e]$ satisfies: (a) it is charged to a distinct degraded episode i (the episode containing the degraded instant of its triggering segment, distinct across intervals by Theorem 3.3(ii)); (b) $s \ge s_i^{\mathrm{ep}}$ and $s \le s_i^{\mathrm{ep}} + 2W$; (c) $e \le \max(e_i^{\mathrm{ep}}, s) + 2W$; hence (d) $e - s \le d_i + 4W$.*

Proof. (a) is Theorem 3.3(i)–(ii): on G the triggering segment contains a degraded instant, which lies in some episode i, and the injection argument of (ii) shows distinct intervals map to distinct episodes. (b): on G no fallback fires in an all-benign segment, so the firing time s is not before the episode's onset; on A_i, $s \le s_i^{\mathrm{ep}} + 2W$. (c): the recovery gate starts monitoring at s; on B_i it fires within $2W$ of the later of the episode's end and its segment start, so $e \le \max(e_i^{\mathrm{ep}}, s) + 2W$. (d): if $e_i^{\mathrm{ep}} \ge s$ then $e - s \le e_i^{\mathrm{ep}} + 2W - s_i^{\mathrm{ep}} \le d_i + 2W$; otherwise ($s$ after the episode end, possible when $d_i < 2W$ and detection lands late) $e - s \le 2W \le d_i + 4W$ using $s \le s_i^{\mathrm{ep}} + 2W \le e_i^{\mathrm{ep}} + 2W$. Both cases satisfy $e - s \le d_i + 4W$. □

The necessity of the coupling is inherited from Theorem 3.4: with $\alpha_{\mathrm{S}} = \infty$ there is an environment and an ungated trigger making $N_T = \Theta(T)$ false interventions, each charged positive cost and none certified—while each individual ledger of Theorem 3.1 remains formally true (the deterministic ledger's constant even survives, since the drift telescopes). What breaks without the gate is not the risk bound but intervention validity and cost control, which no deterministic ledger can certify. The joint statement is thus strictly stronger than the conjunction of its parts.

D Recovery of Prior Controllers: Statements and Proofs

D.1 Special cases recovered (R1–R3)

Each proposition below restricts the triple $(\mathcal{A}, \tilde{\ell}, \alpha_{\mathrm{S}})$; proofs follow.

Proposition D.1 (R1: ACI). *Take $\mathcal{A} = \mathcal{A}_{\mathrm{C}} = \{\mathrm{widen}\}$, $u \equiv 0$, $\tilde{\ell}_t = \mathbf{1}\{y_t \notin \text{interval}\}$. Then Axiom 2.2 is exactly the (unclipped) ACI recursion of Gibbs & Candès (2021), and Theorem 3.1(a) is their coverage theorem with the same constant; their boundedness of the iterate corresponds to Assumption 2.3 with $[\underline{q}, \bar{q}]$ the effective score range.*

Proposition D.2 (R2: Conformal PID). *Take u_t to be the increment of a saturated integrator plus a scorecaster forecast, both $\mathcal{F}_{t-1}$-measurable with total variation $\le U$. Then this (saturated-integrator, bounded-scorecaster) subfamily of the conformal PID controller of Angelopoulos et al. (2023) satisfies Axiom 2.2, and Theorem 3.1(a) reproduces its coverage guarantee; formalizing the paper's observation that the P-part carries validity while I/D affect only efficiency. The recovery covers the bounded-drift subfamily only: Angelopoulos et al. (2023) also prove validity for unsaturated integrators by a different argument, which our recursion class does not subsume.*

Proposition D.3 (R3: online CRC). *Take $\tilde{\ell}_t = \ell(\lambda_t; y_t) \in [0, 1]$ any bounded risk functional of the served parameter λ_t (monotonicity in λ is not needed for the ledger, only for interpreting λ_t as a nested-set threshold), and let Axiom 2.2 act on λ_t. Then Theorem 3.1(a) yields $|T^{-1}\sum_t \ell_t - \alpha| = O(1/\gamma T)$: the online conformal risk control guarantee (Feldman et al., 2023), with batch CRC as its exchangeable specialization (Remark C.4).*

D.2 ANONYMIZED CONTROLLER RECOVERIES (R4–R6)

Three further recoveries (R4–R6) cover the three anonymized deployment controllers of Section 1.1; stratified debt control, operator orchestration, and e-gated fallback. The manuscripts are under review, so the recoveries sit at footnote level[2] and state the full propositions (Propositions D.4 to D.6) with proofs below; the externally checkable recoveries above, together with the decision-layer recoveries of Prinster et al. (2026); Wang & Dobriban (2026) in Section 4, carry the subsumption claim on their own.

The recoveries R1–R3 target published theorems that any reader can check. R4–R6 target anonymized manuscripts under review; so that these recoveries are externally checkable, we restate (anonymized, notation adapted) the exact guarantees being recovered. Full statements also ship in the released supplementary package.

R4 target (Anonymous, 2027a): *For each stratum g of an $\mathcal{F}_{t-1}$-measurable partition, with per-stratum post-action booking and widening recursion (step γ), the realized per-stratum miscoverage satisfies $|T_g^{-1}\sum_{t:g_t=g}\mathrm{err}_t-\alpha|\le C_g/(\gamma T_g)$ on every trajectory, where C_g depends only on the per-stratum threshold range; the controller's debt statistic is the rectified exponentially-discounted integral of the booked residual* $\mathrm{err}_t-\alpha$. Recovered by Theorem 3.1(a) applied per stratum (Proposition D.4).

R5 target (Anonymous, 2027c): *For a registry of repair-operator subsets with paired shadow statistic (risk difference vs. no-op on the same matured outcome), the certificate that an adopted subset does not increase risk is anytime-valid at level α_S: $\mathbb{P}$(adopt while H_0 holds throughout the monitoring segment) $\le\alpha_S$, uniformly over stopping times.* Recovered by Theorem 3.1(b) with the singleton gate family (Proposition D.5); the manuscript's semi-bandit regret bound is an efficiency statement outside the intervention ledgers.

R6 target (Anonymous, 2027b): *With fallback/recovery gates on the clipped model−baseline loss difference, windowed restarts of length W and spending $\alpha_j\propto 1/j(j+1)$: (i) the probability that a fallback fires in a segment where the model is not degraded throughout is at most the gate budget; (ii) under persistent effect Δ satisfying the detectability condition, fallback fires within $W+O(W)$ steps of onset with probability $\ge 1-\beta$, and ex-post excess loss is bounded by the per-step excess times the delay.* Recovered by Theorem 3.1(b) and Theorem C.1 (Proposition D.6).

The statements of R1–R3 appear above; the deferred recoveries of the anonymized deployment controllers (R4–R6) follow.

Proposition D.4 (R4: stratified debt control)**.** *Take $\mathcal{A}_C$ = per-stratum widening, $\mathcal{A}_S$ = a repair library with $\alpha_S=\infty$ (no gate), and per-stratum booking. Then Theorem 3.1(a) per stratum is the safety guarantee of the coverage-debt controller (Anonymous, 2027a) (in our terms, a booked-proxy budget guarantee), and its debt statistic is the rectified discounted integral of the ledger residual. The controller occupies the deterministic row of Figure 1 only; Theorem 3.4 shows the missing cell cannot be filled without a gate.*

Proposition D.5 (R5: operator orchestration)**.** *Take $\mathcal{A}_S$ = subsets of a repair-operator registry with paired shadow statistic D_t^a = risk difference against no-op. Then the anytime e-certificate of Anonymous (2027c) is an instance of Axiom 2.5 and its guarantee is Theorem 3.1(b) for the singleton family; its coverage backstop is Theorem 3.1(a). (Its semi-bandit regret bound concerns efficiency, orthogonal to the intervention ledgers.)*

Proposition D.6 (R6: e-fallback)**.** *Take $\mathcal{A}_S=\{\text{fallback},\text{recover}\}$ with windowed restarts, each gate receiving budget $\alpha_S/2$ with per-segment spending $\alpha_j=(\alpha_S/2)/j(j+1)$ (so the joint spend over both gates is $\le\alpha_S$). Then Theorem 3.1(b) is the false-fallback guarantee of Anonymous (2027b) and Theorem C.1 its detection-delay and ex-post excess-loss bounds; recovery is the reversed null in the same family.*

Proof of Proposition D.1. With $\mathcal{A}=\{\text{widen}\}$ there are no structural interventions, so Axiom 2.4 holds trivially (each step delivered and booked once). With $u\equiv 0$ ($U=u_{\max}=0$) and $\tilde{\ell}_t=\mathrm{err}_t$

[2] R4: per-stratum Theorem 3.1(a) with ungated repairs is the debt controller's safety guarantee (deterministic row of Figure 1 only). R5: the orchestration e-certificate is Axiom 2.5 with a singleton family, its guarantee Theorem 3.1(b). R6: two reversed-null gates with windowed restarts give the false-fallback, delay, and excess-loss bounds via Theorem 3.1(b) and Theorem C.1.

the recursion of Axiom 2.2 is $q_{t+1} = q_t + \gamma(\mathrm{err}_t - \alpha)$: ACI's unclipped update on the served level. When scores are bounded; Assumption 2.3 holds with $[\underline{q}, \bar{q}]$ the score range, and Theorem 3.1(a) gives $|T^{-1}\sum \mathrm{err}_t - \alpha| \le (\bar{q} - \underline{q} + 2\gamma)/(\gamma T)$, the bound of Gibbs & Candès (2021) with their interval length constant. □

Proof of Proposition D.2. The conformal PID update is $q_{t+1} = q_t + \gamma(\mathrm{err}_t - \alpha) + (r_t(\sum_{s\le t}(\mathrm{err}_s - \alpha)) - r_{t-1}(\cdot)) + (\hat{s}_{t+1} - \hat{s}_t)$ where r_t is a saturation function and $\hat{s}$ the scorecaster. Both correction terms are $\mathcal{F}_{t-1}$-measurable at the time they enter (the integrator uses matured errors, the scorecaster is a forecast), and both telescope: the saturated integrator is bounded by its saturation constant and the scorecaster path by its range, so $|\sum_t u_t| \le U$ with U = saturation constant + scorecaster range. Axiom 2.2 holds; Theorem 3.1(a) reproduces the PID coverage bound with the $2U/\gamma T$ term carrying the I/D modules, which affect the constant, never the rate, formalizing the P-carries-validity observation of Angelopoulos et al. (2023). □

Proof of Proposition D.3. The ledger needs only $\ell \in [0, 1]$, which satisfies the proxy boundedness of Axiom 2.2; monotonicity of $\ell(\lambda; y)$ in λ is not used by Theorem 3.1(a); it matters only for interpreting the recursion state λ_t as the threshold of a nested prediction-set family (the served-parameter semantics of risk control), and for Assumption 2.3: for a monotone risk with $\ell(\lambda_{\max}; \cdot) = 0$ and $\ell(\lambda_{\min}; \cdot) = 1$, boundary feasibility holds with $[\underline{q}, \bar{q}] = [\lambda_{\min}, \lambda_{\max}]$. Theorem 3.1(a) then yields $|T^{-1}\sum_t \ell(\lambda_t; y_t) - \alpha| \le (\lambda_{\max} - \lambda_{\min} + 2\gamma)/(\gamma T)$, the online CRC guarantee (Feldman et al., 2023). Batch CRC (Angelopoulos et al., 2024) is the exchangeable, fixed-$\hat{\lambda}$ specialization; we do not claim its finite-sample expectation bound here (Remark C.4). □

Proof of Proposition D.4. Per-stratum booking with post-action accounting is precisely Axiom 2.4 per stratum, and the stratified widening recursion is Axiom 2.2 per stratum, so the per-stratum ledger of Theorem 3.1(a) applies with horizon T_g: the debt controller's safety guarantee, i.e. a booked-proxy budget guarantee. Its debt statistic $D_t^g = (\text{target} - \text{discounted coverage})_+$ is the rectified exponentially-discounted integral of the per-stratum ledger residual $\tilde{\ell}_s - \alpha$, i.e. a monitor of the transient term the ledger bounds. With $\alpha_{\mathrm{S}} = \infty$ its repairs are ungated, so Theorem 3.1(b) is vacuous for it: it occupies only the deterministic row. □

Proof of Proposition D.5. The orchestration certificate tests H_0: the operator subset does not reduce risk against no-op, using the paired difference on the same matured outcome; exactly the shadow statistic D_t^a of Axiom 2.5 with the singleton family $\{a\}$; its anytime validity claim is Theorem 3.1(b) with one gate. Its coverage backstop (per-action conformal widths adapted post-hoc) is an instance of Theorem 3.1(a) with per-action booking. The semi-bandit regret bound governs how quickly the policy finds cost-effective subsets: an efficiency statement about $\sum_t c(a_t)$, complementary to and not implied by the intervention ledgers. □

Proof of Proposition D.6. Fallback and recovery are two gates with reversed nulls on the same paired statistic (D_t = clipped loss difference model − baseline). Windowed restarts of length W and per-gate budget $\alpha_{\mathrm{S}}/2$ with segment spending $\alpha_j = (\alpha_{\mathrm{S}}/2)/j(j+1)$ (so $\sum \alpha_j \le \alpha_{\mathrm{S}}$ jointly over both gates) instantiate Axiom 2.5; Theorem 3.1(b) is then the false-fallback-rate guarantee and Theorem C.1 instantiated with these (α_j, W) is the detection-delay and ex-post excess-loss pair of Anonymous (2027b). □

E Failure Boundaries: Statements and Proofs

Three theorems delimit the axioms: the headline is in Section 3.5, the supporting two are below, and all are proved in this appendix. Theorem E.1 (single booking): booking a step against the pre-intervention output while the user receives the post-intervention one admits delivered risk a constant above budget forever. Theorem 3.4 (no free lunch): with $\alpha_{\mathrm{S}} = \infty$, no trigger rule bounds false-intervention frequency and missed-detection delay simultaneously; the deterministic ledger holds throughout and certifies nothing about intervention validity. Theorem E.2 (wealth debt): a gate without restarts accumulates log-wealth debt through a benign period of length t^* and pays $\Omega(t^*)$ delay after onset; the windowed restarts of Axiom 2.5 cap the debt at $O(W)$.

E.1 FAILURE-BOUNDARY THEOREMS

Of the three failure-boundary theorems, Theorem 3.4 is stated in the main text (Section 3.5); Theorems E.1 and E.2 are stated here; for the mixture-gate case of Theorem E.2, the constant depends on the grid through the smallest surviving component $\underline{\lambda}$. Proofs follow.

Theorem E.1 (Single booking is necessary)**.** *There exist a score sequence and a history-dependent structural policy such that, if a step's proxy is booked against the pre-intervention output while the user receives the post-intervention output (violating Axiom 2.4 while Axiom 2.2 still holds), the realized delivered risk satisfies* $T^{-1}\sum_t \tilde{\ell}_t^{\mathrm{delivered}} \geq \alpha + c$ *for a constant* $c > 0$ *and all sufficiently large* T*.*

Proof of Theorem E.1. Construction. Let scores be i.i.d. $\mathrm{Unif}[0,1]$ and the interval be $[0, q_t]$, so the delivered miscoverage at threshold q is $1-q$. Let the structural library contain a "repair" that replaces the delivered output by one whose score is the original score inflated by a factor $\kappa > 1$ (repair degrades this stream: e.g. smoothing genuine signal), and let the adversarial policy fire the repair at every step. Violate Axiom 2.4 as stated: book $\tilde{\ell}_t^{\mathrm{booked}} = \mathbf{1}\{u_t > q_t\}$ (pre-repair) while delivering the repaired output with miscoverage $\mathbf{1}\{\kappa u_t > q_t\} = \mathbf{1}\{u_t > q_t/\kappa\}$. The recursion sees the booked sequence and, by Theorem 3.1(a) applied to it, steers $T^{-1}\sum \tilde{\ell}_t^{\mathrm{booked}} \to \alpha$, i.e. $q_t \to 1-\alpha$ in Cesàro mean. But the delivered miscoverage at $q = 1-\alpha$ is $\mathbb{P}(u > (1-\alpha)/\kappa) = 1-(1-\alpha)/\kappa = \alpha + (1-1/\kappa)(1-\alpha)$. Take $c = \frac{1}{2}(1-1/\kappa)(1-\alpha) > 0$: for all large T the delivered risk exceeds $\alpha + c$. The same construction covers double booking (book both pre- and post-repair proxies) by halving the effective γ on the delivered sequence and mis-centering the recursion. □

Proof of Theorem 3.4. Fix any trigger rule τ (a sequence of $\mathcal{F}_{t-1}$-measurable firing decisions) with $\alpha_{\mathrm{S}} = \infty$. Consider the environment $\mathcal{E}_0$ in which all pipeline variants have identical conditional loss law (every structural intervention is useless: H_0^a holds always, and firing costs $c(a) > 0$). Two cases, split on the benign firing rate rather than on τ's memory.

(i) Benign firing rate bounded away from zero. If $\liminf_T \frac{1}{T}\sum_{t\leq T} \mathbf{1}\{\tau \text{ fires at } t\} \geq \rho$ for some $\rho > 0$ on $\mathcal{E}_0$, then every firing on $\mathcal{E}_0$ is a false intervention at cost $c(a) > 0$: the false-intervention frequency stays bounded below by ρ, no uniform bound below ρ holds, and cumulative cost grows linearly. This branch already covers every fixed-threshold or finite-memory rule (a sliding-window detector, a CUSUM with a time-invariant threshold): under the benign law such a rule's firing rate is a positive constant, so it cannot satisfy the false-intervention side at any level below that constant.

(ii) Vanishing benign firing rate. Otherwise the firing frequency on $\mathcal{E}_0$ has $\liminf_T = 0$, and the inter-firing gaps on $\mathcal{E}_0$ are unbounded: fewer than δT firings on $[0,T]$ leave some gap of length $\geq 1/\delta - o(1)$, and δ is arbitrarily small along the $\liminf$ subsequence. Fix any L. Choose t_0 at the start of a benign quiet gap of length $\geq 2L$, and let $\mathcal{E}_1$ coincide with $\mathcal{E}_0$ up to t_0 and then degrade the served pipeline by a margin $\delta' > 0$ — large enough that intervention is warranted (the incumbent's conditional risk strictly exceeds budget while a better variant exists) yet small enough that the observable joint law on $[t_0, t_0+L)$ stays within total variation ϵ of the benign law, possible for every finite L as $\delta' \to 0$ since the observable is a bounded statistic whose per-step law varies continuously in the degradation margin (for a W-memory rule each window's law is within $O(W\delta')$ of benign). On that stretch τ's per-step firing probability exceeds its benign rate by at most the total-variation slack, so with probability bounded away from zero it does not fire at all on $[t_0, t_0+L)$: the warranted intervention is missed for L steps, and L was arbitrary.

Hence for every $\epsilon > 0$ no pair (uniform false-intervention bound, uniform missed-detection delay bound) can hold simultaneously over environments. This is the impossibility whose quantified form is the detectability frontier (†) of Theorem C.1 and the wealth debt of Theorem E.2: small margins are where detection delays grow without bound. Throughout; Theorem 3.1(a) applies to the delivered sequence unchanged, since it is trajectory-wise and indifferent to which (useless) variant was delivered: the deterministic ledger certifies nothing about the validity of structural triggers. □

Theorem E.2 (Wealth debt of a restart-free gate)**.** *Consider any e-process gate without restarts whose predictable bets satisfy* $\lambda_s \geq \underline{\lambda} > 0$ *on at least a fraction* $\rho > 0$ *of the benign steps (a gate whose bets vanish trivially has unbounded post-onset delay; for a mixture gate, the conclusion holds with* $\underline{\lambda}$ *the smallest surviving grid component). If we run it through a benign period of length* t^*

during which $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] \le -\mu < 0$*, the gate has detection delay* $\Omega(\rho\,\underline{\lambda}\,\mu\,t^*)$ *after onset with probability* $1 - o(1)$*: the accumulated log-wealth debt must be repaid before the threshold can be crossed. Windowed restarts (Axiom 2.5) cap this debt at* $O(W)$*.*

Proof of Theorem E.2. Let $E_t = \prod_{s\le t}(1+\lambda_s D_s)$ be an e-process gate without restarts, $\lambda_s \in [0, 1/(2B_D)]$ predictable, satisfying the sensitivity condition of the statement: $\lambda_s \ge \underline{\lambda} > 0$ on at least a fraction $\rho > 0$ of the benign steps. During the benign period, $\mathbb{E}[\log(1+\lambda_s D_s) \mid \mathcal{F}_{s-1}] \le \lambda_s \mathbb{E}[D_s \mid \mathcal{F}_{s-1}] \le -\lambda_s\mu \le 0$; the steps with $\lambda_s \ge \underline{\lambda}$ contribute conditional log-drift $\le -\underline{\lambda}\mu$ each, and there are $\ge \rho\,t^*$ of them, while the remaining steps contribute ≤ 0. The log-increments are bounded: $|\log(1+\lambda_s D_s)| \le \log 2$ since $|\lambda_s D_s| \le 1/2$. Azuma–Hoeffding gives, with probability $\ge 1-\beta$,

$$\log E_{t^*} \;\le\; \sum_{s\le t^*} \mathbb{E}[\log(1+\lambda_s D_s) \mid \mathcal{F}_{s-1}] \;+\; 2\log 2 \cdot \sqrt{\tfrac{t^*}{2}\,\ln(1/\beta)}$$

$$\le\; -\rho\,\underline{\lambda}\,\mu\,t^* \;+\; O\big(\sqrt{t^*\log(1/\beta)}\big),$$

and absorbing the deviation term into the linear drift (for $t^* \gg \log(1/\beta)/(\rho\,\underline{\lambda}\,\mu)^2$), $\log E_{t^*} \le -\tfrac{1}{2}\rho\,\underline{\lambda}\,\mu\,t^*$: the wealth enters the onset with log-debt $\Omega(\rho\,\underline{\lambda}\,\mu\,t^*)$. (For a mixture gate, the mixture wealth is at most the maximum component wealth times the grid size, so the same debt holds up to a $\log K$ additive term, with $\underline{\lambda}$ the smallest surviving grid component: the constant depends on the grid, as stated.) Post-onset the log-wealth grows at most $\log(1+\lambda_s D_s) \le \lambda_s D_s \le 1/2$ per step, so crossing any fixed threshold $\log(1/\alpha)$ from $-\Omega(t^*)$ requires $\Omega(t^*)$ steps. With windowed restarts the wealth at the last pre-onset restart is reset to 1, capping the debt at the within-window drift $O(W)$. □

F The Decision Layer: Statements and Proofs

The remaining decision-layer results are stated and proved in full in this appendix. *Recovered objectives and failure boundaries:* the ODM-PS (optimal decision-making based on prediction sets) minimax interface, conformal policy control, and direct robustness control are projections R7–R9 of Theorem 4.2 (Wang & Dobriban, 2026; Prinster et al., 2026; Hu et al., 2026) (Propositions F.4 to F.6); unbounded loss breaks the bridge (F4; Proposition F.7) and *identical* coverage ledgers can differ by $\Omega(1)$ in decision loss (F5; Proposition F.8); the layer is not a corollary of coverage. *Day-level events:* under a two-sided saturation condition; Theorem F.10 gives the day-level violation frequency its own deterministic $O(1/\gamma_D D)$ ledger, step-level ledgers holding unchanged underneath. *The meta layer:* under no-performativity (Assumption G.1) and per-ledger booking (A2′), interleaved Hedge tracks the best fixed library member at anytime-valid rate $O\big(B\sqrt{(H+1)\ln K}/\sqrt{T}\big)$ (Theorem G.2); its switches are ungated, and the *gate-confirmed leader* (GCL) closes that gap; every adopted switch certified at level α_{meta}, switching regret $O(BSW/T)$ on the (Δ, W)-detectable class (Theorem G.3), an intrinsic per-switch price $\Omega(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta)$ (Proposition G.4), matched gap-adaptively (Corollary G.5) and closed in the worst case on gap-homogeneous and c-level-stable classes (Theorems G.6 and G.7); fully unstable gap profiles remain open (the most promising route and its breaking point are below). *The served trajectory, closed:* every executed intervention sits under one joint anytime-valid budget $\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}}$ (Theorem G.8; $K\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}}$ under the deployment-budget convention, the $\ln K$ shift of (†) the quantified power price). No-performativity itself is priced: under ε-performativity counterfactual readings inflate by $k_{\mathrm{ref}}E_T$ (Proposition G.10); E_T is not nonparametrically estimable; the available empirical access is a one-sided served/shadow falsification audit.

F.1 Decision-layer corollaries

Corollary F.1 (Coupled decision ledger)**.** *Under the coupled protocol of Theorem 3.3, with probability at least* $1-\alpha_{\mathrm{S}}$*, simultaneously for all* T*,*

$$\frac{1}{T}\sum_{t\le T}\ell^{\mathrm{dec}}(z_t, y_t) \le \frac{1}{T}\sum_{t\le T} R_t$$

$$+\, B_{\mathrm{dec}}\Big[\alpha + \tfrac{\bar{q}-\underline{q}+2(\gamma+u_{\max}+\kappa)+5(U_0+\kappa)}{\gamma T}\Big].$$

Corollary F.2 (Interventional decision regret)**.** *If $\mathcal{D}$ is the minimax map $\mathcal{D}_\alpha(C) \in \arg\min_{z\in\mathbb{A}} L_C(z;\alpha)$ (well-defined when the* arg min *exists, e.g. compact $\mathbb{A}$ and continuous ℓ^{dec}), then pathwise*

$$\sum_{t\le T} \ell^{\mathrm{dec}}(z_t, y_t) - \sum_{t\le T} \min_{z\in\mathbb{A}} L_{C_t}(z;\alpha) \le B_{\mathrm{dec}}\Big(\alpha T + \tfrac{|q_{T+1}-q_1|+U}{\gamma}\Big):$$

realized decision loss exceeds the best certificate available at each served set by at most the budget allowance: the sequential, intervention-robust analogue of the fixed-set minimax optimality of Wang & Dobriban (2026, Thm. 2.2).

Corollary F.3 (Width is decision cost)**.** *For the newsvendor task ($\mathbb{A} = \mathbb{R}_{\ge 0}$, $\ell^{\mathrm{dec}}(z,y) = h(z-y)_+ + b(y-z)_+$ truncated at B_{dec}) with interval sets, the minimax certificate satisfies $\min_z L_C(z;\alpha) \le \frac{hb}{h+b}\,|C| + \alpha B_{\mathrm{dec}}$ (the ρ-quantile point of the interval, $\rho = b/(h+b)$, attains the in-set term): the width-dependent part of the certificate in Theorem 4.2 is* linear in served width*, and the robust tail contributes only the width-independent constant αB_{dec}. The compliance-at-width frontier of Appendix H.1 therefore converts, pound for pound, into a decision-loss frontier; measured in Appendix H.10.*

F.2 Recovered decision objectives

Proposition F.4 (R7: ODM-PS)**.** *Take the static SIP ($\mathcal{A}_{\mathrm{S}} = \emptyset$, no structural intervention, fixed split-conformal set C with exchangeable calibration) and $\mathcal{D} = \mathcal{D}_\alpha$. Then Theorem 4.2(i) plus marginal coverage yields $\mathbb{E}[\ell^{\mathrm{dec}}(z,Y)] \le L_C(\mathcal{D}_\alpha(C);\alpha) + B_{\mathrm{dec}}\,\alpha$-type control with the minimax action of Wang & Dobriban (2026, Lem. 2.1, Thm. 2.2): their decision rule and its risk certificate are the trivial-library projection of Theorem 4.2. We recover the minimax interface,* not *their oracle set-design theorem (the duality characterization of optimal coverage allocation across inputs), which optimizes over the set constructor: an axis the SIP holds fixed.*

Proposition F.5 (R8: conformal policy control)**.** *Identify the policy space of Prinster et al. (2026) with the action space $\mathbb{A}$, the safe reference policy π_0 with the fallback pipeline, and proposed policy updates with structural actuators in $\mathcal{A}_{\mathrm{S}}$. CPC's likelihood-ratio clipping is a calibration-actuator parameter (it tunes how far the served object may move from the reference), and its weighted-conformal risk calibration is an instance of the Axiom 2.5 gate: an update is adopted only when a statistic certifies the reference risk level is preserved. Under this identification, CPC's guarantee (deviation from the safe policy's risk level is controlled at each adopted update) is the decision ledger of Theorem 4.2 restricted to policy-valued actions with R_t the reference-policy certificate. The correspondence is structural: CPC's finite-sample constants rest on exchangeability-based weighted conformal calibration and stability remainders that the adversarial SIP setting does not (and cannot) reproduce; what the SIP adds is survival of the guarantee under history-dependent switching among updates.*

Proposition F.6 (R9: direct robustness control)**.** *Conformal robustness control (CRC-Robust; Hu et al., 2026) observes that coverage is sufficient but not necessary for downstream robustness: their target is the robustness constraint $\mathbb{P}\{\phi(Y, z(X)) \le r(X)\} \ge 1-\alpha$ on decision loss ϕ and served certificate r, not the coverage condition $\mathbb{P}\{Y \in U(X)\} \ge 1-\alpha$ that implies it. In the restriction triple $(\mathcal{A}, \tilde{\ell}, \alpha_{\mathrm{S}})$, keep $\mathcal{A}$ and α_{S} arbitrary and restrict only the ledger functional to the decision-level violation indicator, $\tilde{\ell}_t = \mathbf{1}\{\ell^{\mathrm{dec}}(z_t, y_t) > R_t\}$, which Axiom 2.2 accepts unchanged (a $[0,1]$ proxy of the delivered output). Then Theorem 3.1(a) controls the realized frequency of certificate violations at $|T^{-1}\sum_t \tilde{\ell}_t - \alpha| = O(1/\gamma T)$ on every trajectory; the sequential, intervention-robust counterpart of their i.i.d.-based test-time calibration guarantee ($\mathbb{P}\{\phi(Y_{n+1}, z(X_{n+1})) \le r(X_{n+1})\} \ge 1-\alpha$), with the same target level and without the $B_{\mathrm{dec}}\,\alpha$ overhead of the coverage route. The recovery is of the constraint functional only, and the delimitation is sharp in both directions: their contribution is the* optimization *of the static set construction under this constraint (batch training with non-asymptotic robustness and optimality guarantees), an axis the SIP holds fixed; as R7 recovers ODM-PS's minimax interface but not its oracle set-design theorem; conversely, their guarantees are i.i.d.-based and static, and do not survive history-dependent structural interventions, which is what the SIP ledgers add.*

F.3 DECISION-LEVEL FAILURE BOUNDARIES

Proposition F.7 (F4: unbounded loss breaks the bridge)**.** *For every $\alpha \in (0,1)$ and every coverage-controlling policy there is a task with $B_{\mathrm{dec}} = \infty$ (e.g. newsvendor with unbounded shortage penalty) and a data law under which $T^{-1}\sum_t \tilde{\ell}_t \to \alpha$ while $T^{-1}\sum_t \ell^{\mathrm{dec}}(z_t, y_t) = \infty$ for every T: no coverage ledger controls expected decision loss without a loss bound (or a hedging action whose worst case is finite; the practical fix, priced by exactly the B_{dec} that reappears in Theorem 4.2).*

Proposition F.8 (F5: equal coverage, $\Omega(1)$ decision gap)**.** *There exist two policies with identical realized coverage ledgers ($\tilde{\ell}$-sequences equal step by step) and an asymmetric decision task on which their average decision losses differ by a constant independent of T: interval* placement*, invisible to the coverage ledger, is decision-relevant. It is invisible to the certificate as well: the interval minimax certificate prices served width only (Corollary F.3), so the two policies' certified terms coincide and the $\Omega(1)$ gap appears only in realized decision loss—the certified term bounds realized loss but is not a surrogate for it.*

Proofs below. Together: coverage buys decision control only through the bridge (bounded loss, priced allowance), and the served certificate R_t ranks policies by width where coverage is silent, while placement differences escape both ledgers and surface only in realized loss. This is the theory behind the decision experiments of Appendix H.10.

F.4 A DAY-LEVEL JOINT LEDGER

The storage task exposes a gap that the per-step machinery cannot close: the certified daily profit floor is an event on *two* coordinates of the same day, while the served bands are calibrated marginally per step; a per-step ledger at level α says nothing quantitative about the frequency of day-level floor violations. Rather than intersecting marginal bands (a Bonferroni-style union bound that inflates with the number of coordinates and still yields only a per-day probability under distributional assumptions we do not make), we control the day-level event by the same device that controls the step-level one: give the joint event its own budget recursion, on its own timescale.

Partition time into blocks (days) B_d of length K, $d = 1, \dots, D$. A *block decision rule* maintains a scalar inflation parameter $\lambda_d \in \mathbb{R}$; at the start of block d it computes a block action and an observable certificate $\Pi_d(\lambda_d)$ from the served sets of block d inflated by λ_d (for storage: schedule and floor from $[L_t - \lambda_d,\, U_t + \lambda_d]$), and at block end the joint miss indicator $m_d \in \{0,1\}$ (for storage: realized profit below the certified floor) matures and drives

$$\lambda_{d+1} = \lambda_d + \gamma_D\,(m_d - \alpha_D), \qquad \gamma_D > 0. \tag{2}$$

Assumption F.9 (Two-sided saturation)**.** There exist finite $\underline{\lambda} \le \bar{\lambda}$ such that for every block d and every history, $\lambda_d \ge \bar{\lambda}$ implies $m_d = 0$ and $\lambda_d \le \underline{\lambda}$ implies $m_d = 1$.

For the storage instantiation on a bounded stream; Assumption F.9 holds automatically *whatever the schedule does*: the certified floor $\max_{b<s}\,\eta(L_s - \lambda) - (U_b + \lambda)$ is decreasing in λ and diverges to $-\infty$ (every entry of the max does), while realized profit is bounded on a bounded stream, so large λ forces $m_d = 0$ and very negative λ forces $m_d = 1$, for *every* choice of charge/discharge hours. For this particular certificate the inflation happens to enter as the schedule-independent offset $-(1+\eta)\lambda$, so the argmax schedule is unchanged and m_d is even monotone in λ_d; the theorem does not use this, and the saturation route is what carries the guarantee to block rules where the inflation *does* interact with the schedule (per-hour or asymmetric inflation, multi-cycle schedules), where no per-schedule monotonicity is available.

Theorem F.10 (Two-timescale joint ledger)**.** *Run the step-level SIP unchanged (Axioms 2.2, 2.4 and 2.5 on the served sets) and the block recursion Equation* (2) *on top, initialized at $\lambda_1 \in [\underline{\lambda}, \bar{\lambda}]$ (e.g. $\lambda_1 = 0$ after enlarging the saturation interval to contain it, which Assumption F.9 permits), with Assumption F.9. Then, on every trajectory, simultaneously: (i) all step-level ledgers of Theorems 3.1 and 3.3 hold unchanged; the inflation acts inside the decision interface only and never touches a served set; (ii) $\lambda_d \in [\underline{\lambda} - \gamma_D,\, \bar{\lambda} + \gamma_D]$ for every index $d \ge 1$ at which the recursion defines λ_d; and (iii) for every $D \ge 1$,*

$$\Big|\frac{1}{D}\sum_{d=1}^{D} m_d - \alpha_D\Big| \;\le\; \frac{\bar{\lambda} - \underline{\lambda} + 2\gamma_D}{\gamma_D\, D}.$$

The bound is deterministic and holds for arbitrary history-dependent interventions upstream: structural shocks move the served bands, and the block recursion self-corrects on every trajectory.

The guarantee has exactly the epistemic status of Theorem 3.1(a) (a long-run realized-frequency ledger, not a per-day or distributional claim) but it is a *joint* ledger: the controlled quantity is the day-level event that the marginal bands could not speak about. The price is observable ex ante as the floor tightness lost to the inflation λ_d, and is measured in Appendix H.10. Proof below.

Proof of Theorem 4.2(i) (bridge inequality). Fix t. If $y_t \in C_t$, then $\ell^{\mathrm{dec}}(z_t, y_t) \le \sup_{y \in C_t} \ell^{\mathrm{dec}}(z_t, y) = \ell_{C_t}(z_t) \le L_{C_t}(z_t; \alpha)$, since the second term of L is nonnegative. If $y_t \notin C_t$, then $\ell^{\mathrm{dec}}(z_t, y_t) \le B_{\mathrm{dec}} = B_{\mathrm{dec}} \mathbf{1}\{y_t \notin C_t\}$, and $R_t \ge 0$. Either way the claimed inequality holds; no distributional assumption and no property of the intervention policy was used: C_t and z_t may depend on the entire history in any measurable way. □

Proof of Theorem 4.2(ii) (decision ledger). Sum the bridge inequality over $t \le T$ and divide by T:

$$\frac{1}{T}\sum_t \ell^{\mathrm{dec}}(z_t, y_t) \le \frac{1}{T}\sum_t R_t + B_{\mathrm{dec}} \cdot \frac{1}{T}\sum_t \tilde{\ell}_t, \qquad \tilde{\ell}_t = \mathbf{1}\{y_t \notin C_t\}.$$

The miscoverage indicator of the *delivered* set is a valid $[0,1]$ risk proxy under Axiom 2.4 (single booking), so Theorem 3.1(a) applies directly: $\frac{1}{T}\sum_t \tilde{\ell}_t \le \alpha + \frac{|q_{T+1} - q_1| + U}{\gamma T}$ on every trajectory, and under Assumption 2.3 the range term is bounded uniformly in T. For the coupled statement, replace the invocation of Theorem 3.1(a) by Theorem 3.3: on the Ville good event (probability $\ge 1 - \alpha_{\mathrm{S}}$, uniformly over stopping times) the same average is bounded with the constant $\bar{q} - \underline{q} + 2(\gamma + u_{\max} + \kappa) + 5(U_0 + \kappa)$, and multiplying by B_{dec} preserves the bound. No step used independence, exchangeability, or a fixed pipeline. □

Proof of Corollary F.2. With $\mathcal{D} = \mathcal{D}_\alpha$, $R_t = L_{C_t}(\mathcal{D}_\alpha(C_t); \alpha) = \min_z L_{C_t}(z; \alpha)$ by definition of the minimax map. Subtract $\sum_t \min_a L_{C_t}(a; \alpha)$ from both sides of the summed bridge inequality and apply Theorem 3.1(a) to the indicator sum as above. □

Proof of Corollary F.3. For an interval $C = [L, U]$ and the (truncated) newsvendor loss, order at the interior point $z^* = L + \rho(U - L)$ with $\rho = b/(h + b)$. For $y \in C$: overage is at most $h(z^* - L) = h\rho|C|$ and underage at most $b(U - z^*) = b(1 - \rho)|C|$; both equal $\frac{hb}{h+b}|C|$, so $\ell_C(z^*) \le \frac{hb}{h+b}|C|$. The second term of $L_C(z^*; \alpha)$ adds $\alpha(\bar{\ell}(z^*) - \ell_C(z^*))_+ \le \alpha B_{\mathrm{dec}}$, a constant that does not scale with $|C|$; for the width-scaling claim we use the certified in-set term, which is linear in $|C|$ as stated. Hence $\min_z L_C(z; \alpha) \le L_C(z^*; \alpha) \le \frac{hb}{h+b}|C| + \alpha B_{\mathrm{dec}}$. □

Proof of Theorem F.10. (i) is by construction: the block rule reads the served sets and writes only to the decision interface (the pair block action / certificate); no quantity entering Axioms 2.2, 2.4 and 2.5 is modified, so every step-level statement holds verbatim on the same trajectory.

(ii) Confinement. Suppose $\lambda_d > \bar{\lambda} + \gamma_D$ for some d, and take the first such d; $d \ge 2$ since $\lambda_1 \in [\underline{\lambda}, \bar{\lambda}]$. Since single-step moves of Equation (2) are bounded by $\gamma_D \max(\alpha_D, 1 - \alpha_D) \le \gamma_D$, we have $\lambda_{d-1} \ge \lambda_d - \gamma_D > \bar{\lambda}$, so Assumption F.9 gives $m_{d-1} = 0$ and $\lambda_d = \lambda_{d-1} - \gamma_D \alpha_D \le \lambda_{d-1}$; by minimality of d, $\lambda_{d-1} \le \bar{\lambda} + \gamma_D$, so $\lambda_d \le \bar{\lambda} + \gamma_D$, a contradiction. The lower bound is symmetric: $\lambda_{d-1} < \underline{\lambda}$ forces $m_{d-1} = 1$ and an upward move of size $\gamma_D(1 - \alpha_D) \ge 0$, so $\lambda_d \ge \lambda_{d-1} \ge \underline{\lambda} - \gamma_D$ by minimality. The argument uses only Equation (2) and λ_1, so it covers every index at which the recursion defines λ, including $D + 1$.

(iii) Telescoping Equation (2): $\sum_{d=1}^{D}(m_d - \alpha_D) = (\lambda_{D+1} - \lambda_1)/\gamma_D$. By (ii) applied at indices 1 and $D + 1$, both lie in $[\underline{\lambda} - \gamma_D, \bar{\lambda} + \gamma_D]$, so $|\lambda_{D+1} - \lambda_1| \le \bar{\lambda} - \underline{\lambda} + 2\gamma_D$. Divide by D. No step used any property of the data, the schedule map $\lambda \mapsto$ (block action), or the upstream intervention policy: m_d may depend on the entire history arbitrarily, exactly as $\tilde{\ell}_t$ may in Theorem 3.1(a).

Storage saturation constants: on a stream with $|y_t| \le M$ and served bands of width at most $w_{\max}$ (finite by Assumption 2.3 and the boundedness of the score range), realized profit lies in $[-(1 + \eta)M, (1 + \eta)M]$ while the certified floor is at most $\eta(M + w_{\max} - \lambda) + M + w_{\max} - \lambda$

at inflation λ and at least $-\eta(M + w_{\max}) - M - w_{\max} + (1 + \eta)|\lambda|$ for $\lambda < 0$; solving floor $\leq -(1 + \eta)M$ (which forces $m_d = 0$, since realized profit is never below the floor) and floor $> (1 + \eta)M$ (which forces $m_d = 1$) gives finite $\bar{\lambda}, \underline{\lambda}$ depending only on $(M, w_{\max}, \eta)$. □

Proof of Proposition F.4. With $\mathcal{A}_{\mathrm{S}} = \emptyset$ there are no structural interventions and no shocks ($u_t \equiv 0$); with a fixed exchangeably-calibrated set C the miscoverage indicator has marginal expectation $\leq \alpha$. Taking expectations in Theorem 4.2(i) gives $\mathbb{E}[\ell^{\mathrm{dec}}(\mathcal{D}_\alpha(C), Y)] \leq L_C(\mathcal{D}_\alpha(C); \alpha) + B_{\mathrm{dec}}\,\alpha$; the fixed-set analysis of Wang & Dobriban (2026) is sharper; it bounds $\mathbb{E}[\ell^{\mathrm{dec}}]$ by $L_C(a; \alpha)$ itself when α upper-bounds the miscoverage probability, since L's second term already prices the out-of-set event at its worst case times α (replace $\bar{\ell}(a)$-truncation by conditioning; the same one-line argument). Their Lemma 2.1 (validity of the certified risk) and Theorem 2.2 (minimaxity of $\mathcal{D}_\alpha$) are then the trivial-library, exchangeable-data instance: the SIP machinery contributes nothing (there is nothing to intervene on), and the recovered statement is theirs. What the SIP adds beyond this projection (control under nontrivial $\mathcal{A}_{\mathrm{S}}$) is Theorem 4.2. We do not recover their oracle set-design duality, which optimizes over the set constructor, held fixed here. □

On the wording of Proposition F.5. The proposition asserts a structural identification, not a theorem-for-theorem recovery; the displayed correspondence (policy space → action space, safe policy → fallback pipeline, clipping → calibration parameter, weighted-conformal certificate → Axiom 2.5 gate) is checked against the published statements of Prinster et al. (2026), and the guarantee that survives in the SIP setting is the one stated: certificate-gated adoption controls deviation from the reference risk level, with the SIP supplying validity under history-dependent switching where CPC's exchangeability-based constants do not apply.

Proof of Proposition F.7. Newsvendor with $h = 1$ and unbounded underage: let $\ell^{\mathrm{dec}}(z, y) = (z - y)_+ + \infty \cdot \mathbf{1}\{y > z + 1\}$ (any loss unbounded on out-of-set outcomes works). Let the data law put mass α on a diverging sequence of outcomes outside every served set: under Axiom 2.2 the served threshold moves by at most γ per step, so every realized set sequence is contained in sets of diameter $O(\gamma t)$ at time t, and outliers $y_t = t^2$ (any super-linear growth suffices) eventually escape every such sequence no matter how the recursion widens; the calibration recursion, which sees only the binary indicator, is unaffected and its ledger converges to α. Every step on which the outlier lands costs $+\infty$, and such steps have frequency $\alpha > 0$: average decision loss is $+\infty$ for every T while the coverage ledger is exact. The practical fix (an action hedging all outcomes to finite worst case) is precisely a finite B_{dec}, and its price is the $B_{\mathrm{dec}}\alpha$ term of Theorem 4.2. □

Proof of Proposition F.8. Let y_t alternate deterministically between 0 and 1, and compare two interval policies of identical width and identical coverage sequence: policy A serves $[y_t - 0.1, y_t + 0.9]$ (covers, low edge tight) and policy B serves $[y_t - 0.9, y_t + 0.1]$ (covers, high edge tight); both have $\tilde{\ell}_t \equiv 0$, identical coverage ledgers, identical widths. Take asymmetric newsvendor loss with $b \gg h$ and $\mathcal{D} = \mathcal{D}_\alpha$, which orders near the upper endpoint: policy A orders $\approx y_t + 0.9 - \epsilon$ and pays overage $\approx 0.9h$ per step; policy B orders $\approx y_t + 0.1 - \epsilon$ and pays $\approx 0.1h$. The per-step gap is $0.8h = \Omega(1)$, uniformly in T. The certified terms, by contrast, coincide: under Corollary F.3 the interval minimax certificate is linear in served width alone (the $\frac{hb}{h+b}|C_t| + \alpha B_{\mathrm{dec}}$ bound), both policies serve width 1.0, so $R_t^{\mathrm{A}} = R_t^{\mathrm{B}}$ step by step. The separation is thus invisible to the coverage ledger *and* to the certified terms; it appears only in the realized decision loss that the certificate bounds. □

G META- AND COMPOSITE-LEDGER GUARANTEES

This section develops the two layers above the intervention library itself: a meta-policy that certifies switching among library members (regret, gated adoption, and matching lower bounds), and the composite ledger that prices the trajectory actually served under one joint budget.

G.1 A META-POLICY OVER THE LIBRARY

Everything so far certifies a *fixed* policy. But the experiments show that which policy is best depends on the scenario (recalibration-only wins when both views are corrupted, the full library wins on asymmetric incidents) and no single deployment knows its scenario in advance. The natural question

is whether one can run a *meta-policy* over the library of SIP policies and provably track the best of them, without surrendering the ledgers. The obstacle is semantic, not technical: a meta-policy serves one output per step, while Axiom 2.4 books each step against the delivered output; a mixture appears to double-book. One assumption, already implicit in every experiment in this paper, dissolves the obstacle.

Assumption G.1 (No performativity)**.** The stream (x_t, y_t) is exogenous: it does not depend on the served outputs. Consequently the library is *fully simulable*: each base policy π^k's trajectory (its candidate sets C_t^k, its calibration recursion, its e-gates) is a measurable function of the shared history alone, computable in silico whether or not π^k was served.

Under Assumption G.1 we adopt the *per-ledger booking semantics* (A2′): each of the K base policies books every step exactly once against *its own* (possibly counterfactual) output, and the meta-policy books every step exactly once against the *served* output. No ledger is double-booked and no step is booked against an output its ledger's owner did not produce, so every base-policy theorem in this paper applies unchanged to the in-silico trajectories; nothing upstream is re-proved or re-run. Let $\ell_t^k \in [0, B]$ be a bounded per-step decision-relevant loss of policy k's output (e.g. the truncated newsvendor loss of Corollary F.3), observable for all k once y_t matures with delay H. The meta-policy is sampled exponential weights (Cesa-Bianchi & Lugosi, 2006) run as $H{+}1$ interleaved instances (one per residue class mod $H{+}1$, so each instance only ever needs matured feedback (Weinberger & Ordentlich, 2002)), instance learning rate $\eta_n = \sqrt{8 \ln K / n}$ at its n-th own step, served policy sampled from the current weights.

Theorem G.2 (Meta-policy regret over the library)**.** *Under Assumption G.1, bounded losses, and A2′, the sampled interleaved-Hedge meta-policy satisfies: for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $T \geq 1$,*

$$\frac{1}{T}\sum_{t\leq T} \ell_t^{\mathrm{meta}} \;\leq\; \min_{k\leq K} \frac{1}{T}\sum_{t\leq T} \ell_t^k \;+\; \frac{B\sqrt{2(H{+}1)\ln K}}{\sqrt{T}} + \frac{(H{+}1)B\sqrt{\ln K/8}}{T} \;+\; \frac{2B\sqrt{\ln \frac{2(j_T+1)(j_T+2)}{\delta}}}{\sqrt{T}},$$

where $j_T = \lceil \log_2(2T) \rceil$. Moreover (i) every ledger of Theorems 3.1, 3.3, 4.2 and F.10 holds for each base policy's in-silico trajectory; and (ii) if some base policy k^* is budget-compliant, then combining with Theorem 4.2 for k^*, the served *average decision loss is within $O(1/\gamma T) + O\big(\sqrt{(H{+}1)\ln(K/\delta)}/\sqrt{T}\big)$ of k^*'s certified level. The regret is to the best* fixed *library member in hindsight, not to a switching sequence.*

One boundary of Theorem G.2: its expert switches are *not* e-gated, so the anytime-valid intervention ledger Theorem 3.1(b) is inherited only by the base policies' in-silico gates; at that meta level, regret *replaces* gating as the validity currency, and the comparator is fixed. The proof below is a composition of the anytime Hedge bound per interleaved instance, Cauchy–Schwarz across instances, and a Ville-type uniform Azuma bound for the sampling deviation with the same doubling-epoch spending pattern as our gates; the SIP content is entirely in Assumption G.1 + A2′. Measured in E6 (Appendix H.11).

G.2 Switching comparators and partial gating

Both boundaries of Theorem G.2 (the fixed comparator and the ungated switches) can be pushed back simultaneously, and the price of doing so can be quantified from below. The protocol separates *proposal* from *adoption*. The *gate-confirmed leader* (GCL) meta-policy nominates, at each step, the *windowed leader*: the expert with the smallest cumulative matured loss over the trailing W matured steps (a sliding window; switch-adaptive by construction, with no learning rate and no weight state, in contrast to ungated tracking algorithms such as Fixed-Share (Herbster & Warmuth, 1998)). The meta-policy *serves* an incumbent expert; if the nominated candidate differs from the incumbent, a paired mixture-betting e-process: the same gate as Axiom 2.5, window W, spending $\alpha_{\mathrm{meta}}/j(j{+}1)$ on its j-th monitoring segment, reset whenever the nominated candidate changes; runs on the matured in-silico loss differential $(\ell_t^{\mathrm{inc}} - \ell_t^{\mathrm{cand}})/B \in [-1, 1]$. Only when the gate fires does the incumbent switch: an *adopted* meta switch.

Call a comparator segment (a maximal run of a comparator sequence σ on one expert k) (Δ, W)-*detectable at level* δ *by horizon* T if throughout the segment $\mathbb{E}[\ell_t^j - \ell_t^k \mid \mathcal{F}_{t-1}] \geq \Delta$ for every $j \neq k$, where Δ satisfies both the nomination margin $\Delta \geq B\sqrt{2\ln(2KT(T+1)/\delta)/W}$ and the grid detectability condition (†) of Theorem C.1 at window W; the meta-level transcription of the base gates' power condition.

Theorem G.3 (Switching regret with partially gated meta-switches)**.** *Under Assumption G.1, bounded losses, and A2′, the GCL meta-policy satisfies:* ***(a) (Adoption validity, anytime.)*** *With probability at most* α_{meta}*, uniformly over stopping times, some adopted switch fires in a monitoring segment throughout which* $\mathbb{E}[\ell_t^{\mathrm{inc}} - \ell_t^{\mathrm{cand}} \mid \mathcal{F}_{t-1}] \leq 0$*: every adopted meta switch carries the same anytime-valid certificate as a base-level structural intervention.* ***(b) (Tracking, detectable class.)*** *For every* $\delta \in (0,1)$*, every horizon* $T \geq 1$*, and every comparator sequence* σ *with* S *switches all of whose segments are* (Δ, W)*-detectable at level* δ *by horizon* T*, with probability at least* $1 - \alpha_{\mathrm{meta}} - \beta_{\mathrm{det}} - \delta$*,*

$$\frac{1}{T}\sum_{t\leq T}\ell_t^{\mathrm{GCL}} \;\leq\; \frac{1}{T}\sum_{t\leq T}\ell_t^{\sigma_t} \;+\; \frac{B\,(S+1)\,\big(2W+2H+2\big)}{T},$$

where β_{det} *is the total gate-power spending of (†). The excess is* $O\big(BSW/T\big)$*: linear in the number of comparator switches, one window of nomination delay plus one window of confirmation delay per switch, and no* $\sqrt{T}$ *term.*

Proposition G.4 (Certification lower bound)**.** *There is a two-expert library with i.i.d. losses in* $[0, B]$ *and per-step gap* $\Delta \in (0, B/4]$ *such that* any *meta-policy whose adoptions are* α_{meta}*-certified in the sense of Theorem G.3(a) suffers expected excess loss at least* $c\,\min\{\Delta T,\ B^2\ln(1/2\alpha_{\mathrm{meta}})/\Delta\}$ *over the best expert, for a universal constant* c*. Taking* $\Delta = B\sqrt{\ln(1/2\alpha_{\mathrm{meta}})/T}$ *shows a worst-case excess of* $\Omega\big(B\sqrt{T\ln(1/\alpha_{\mathrm{meta}})}\big)$*: the per-switch delay cost in Theorem G.3(b) is not an artifact of our gate; certification itself has this price. (The upper bound pays* $O(BW)$ *per switch while the lower bound demands* $\Omega(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta)$*; since (†) forces* $W = \Omega(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta^2)$*, a factor of order* B/Δ *separates them. Corollary G.5 closes this gap* instance-adaptively*: when the served gap is of order* Δ *(which is the regime of this construction) upper and lower bounds match up to constants and a log factor; the* worst-case *factor survives only on instances whose served gap far exceeds* Δ*, where this construction is not binding; Theorem G.6 closes it there too on the gap-homogeneous class, and Theorem G.7 on the level-stable heterogeneous class; only unstable gap profiles remain open.)*

Corollary G.5 (Gap-adaptive tracking)**.** *In the setting of Theorem G.3(b), let* $\bar{g}_m \in [\Delta/B, 1]$ *denote the largest average* stepwise-maximal *normalized conditional gap* $\frac{1}{n}\sum_t \max_{j'\neq k_m}\mathbb{E}[(\ell_t^{j'} - \ell_t^{k_m})/B \mid \mathcal{F}_{t-1}]$ *over initial stretches* $n \leq 2W+2H+2$ *of comparator segment* m *(the maximum over experts is taken* inside *the sum, at each step, because the incumbent identity may change during the stretch; the lower endpoint* Δ/B *is a consequence of the* (Δ, W)*-detectability assumption, not an extra constraint: throughout a detectable segment every stepwise maximum is at least* Δ/B*, hence so is every prefix average). Then for every* $\delta' \in (0,1)$*, with probability at least* $1 - \alpha_{\mathrm{meta}} - \beta_{\mathrm{det}} - \delta - \delta'$*,*

$$\frac{1}{T}\sum_{t\leq T}\ell_t^{\mathrm{GCL}} \;\leq\; \frac{1}{T}\sum_{t\leq T}\ell_t^{\sigma_t} \;+\; \frac{B}{T}\sum_{m=1}^{S+1}\Big(\bar{g}_m\,(2W+2H+2) + \sqrt{2(2W+2H+2)\ln\tfrac{2(S+1)(2W+2H+2)}{\delta'}}\Big).$$

On the instance of Proposition G.4, $\bar{g}_m = \Delta/B$ *and (†) forces* $W = \Theta(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta^2)$*, so the per-switch cost is* $\Theta(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta)$*; matching the lower bound up to constants. The worst-case gap of Proposition G.4 thus lives only on instances with* $B\bar{g}_m \gg \Delta$*; Theorem G.6 below closes it on the gap-homogeneous part of that range with a multi-scale grid, and Theorem G.7 extends the closure to heterogeneous segments whose per-expert gaps are level-stable.*

The instance-adaptive bound still pays the window W sized for the *smallest* detectable gap: a segment whose actual gap is $G \gg \Delta$ waits the full $O(W) = O(B^2\ln(1/\alpha_{\mathrm{meta}})/\Delta^2)$ delay even though a gap G is statistically detectable in $O(B^2\ln(1/\alpha_{\mathrm{meta}})/G^2)$ steps. Running the gate at *all* dyadic scales removes this mismatch. The *multi-scale* GCL (MS-GCL) runs a dyadic grid of windows $W_i = 2^i W_0$, $i = 0, \ldots, I$, with the meta budget split as $\alpha_{\mathrm{meta}}/(I+1)$ per scale; scale

i nominates the trailing-W_i matured leader, and runs a *candidate-indexed* confirmation gate at window W_i on the matured differential $(\ell_t^{\mathrm{inc}_t} - \ell_t^{\mathrm{cand}_i})/B$ against the *current* incumbent inc_t (an $\mathcal{F}_{t-1}$-measurable index, so the null supermartingale argument is unchanged); the gate's segment counter advances on candidate changes, firings, and window expiries, but *not* on incumbent changes, so an intermediate adoption does not reset a running confirmation. When any scale's gate fires, the incumbent becomes that scale's candidate. Call segment m (G_m, W_i)-*detectable* if throughout it $\mathbb{E}[\ell_t^j - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \geq G_m$ for every $j \neq k_m$, with $G_m \geq B\sqrt{2\ln(2K(I{+}1)T(T{+}1)/\delta)/W_i}$ and (†) holding at window W_i and budget $\alpha_{\mathrm{meta}}/(I{+}1)$; write i_m for the smallest such scale ($W_{i_m} = O(B^2\ln(KIT/\delta\alpha_{\mathrm{meta}})/G_m^2)$ when it exists in the grid), and call the segment c-*gap-homogeneous* if additionally $\max_{j\neq k_m}\mathbb{E}[\ell_t^j - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \leq c\,G_m$ throughout.

Theorem G.6 (Multi-scale certified tracking: worst-case closure on the gap-homogeneous class)**.** *Under Assumption G.1, bounded losses, and A2′, MS-GCL satisfies:* ***(a)*** *the adoption-validity certificate of Theorem G.3(a), with the null read against the current incumbent ($\mathbb{E}[\ell_t^{\mathrm{inc}_t} - \ell_t^{\mathrm{cand}} \mid \mathcal{F}_{t-1}] \leq 0$ throughout the monitoring segment), at total level α_{meta}.* ***(b)*** *For every $\delta, \delta' \in (0,1)$ and every comparator σ with S switches whose segment m is (G_m, W_{i_m})-detectable for some grid scale i_m, with probability at least $1 - \alpha_{\mathrm{meta}} - \beta_{\mathrm{det}} - \delta - \delta'$,*

$$\frac{1}{T}\sum_{t\leq T}\ell_t^{\mathrm{MS}} \;\leq\; \frac{1}{T}\sum_{t\leq T}\ell_t^{\sigma_t} \;+\; \frac{B}{T}\sum_{m=1}^{S+1}\Big(\bar{g}_m\,(2W_{i_m}{+}2H{+}2) + \sqrt{2(2W_{i_m}{+}2H{+}2)\ln\tfrac{2(S+1)(2W_I+2H+2)}{\delta'}}\Big),$$

with $\bar{g}_m$ the stepwise-maximal average gap of Corollary G.5 over initial stretches $n \leq 2W_{i_m} + 2H + 2$. In particular, if segment m is c-gap-homogeneous then $B\bar{g}_m \leq c\,G_m$ and its per-switch cost is $O\big(c\,B^2\ln(KIT/\delta\alpha_{\mathrm{meta}})/G_m\big)$: on the c-gap-homogeneous detectable class the worst-case per-switch cost is $O\big(c\,B^2\ln(KIT/\delta\alpha_{\mathrm{meta}})/\Delta\big)$, matching the certification lower bound of Proposition G.4 (whose instance is 1-gap-homogeneous at level $G_m = \Delta$) up to constants, the homogeneity factor c, and logarithmic factors; simultaneously at every gap level $G_m \in [\Delta, B]$*, since Proposition G.4 applied at gap G_m demands $\Omega(B^2\ln(1/\alpha_{\mathrm{meta}})/G_m)$.*

Two remarks on the certificate's semantics. The null in (a) is *incumbent-varying* ("the candidate is never better than whoever is currently incumbent") which is semantically weaker than the fixed-pair null of Theorem G.3(a): if the incumbent changed mid-segment, a certified adoption asserts the candidate failed to dominate a *moving* baseline, not any single policy; this is the price of not resetting on incumbent changes. Symmetrically, the detectability condition (†) in (b) is applied to a differential whose composition changes with the incumbent; this is licensed because (as in the power argument of Theorem C.1) the mixture-betting wealth growth uses only the *uniform* conditional-mean lower bound $\mathbb{E}[\ell_t^j - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \geq G_m$ for all $j \neq k_m$, not the identity of the served j.

Theorem G.6 closes the worst-case B/Δ factor of Proposition G.4 on the gap-homogeneous detectable class: the single-scale bound pays $G_m \cdot B^2/\Delta^2$ per switch where the lower bound demands B^2/G_m, and the dyadic grid replaces the former by the latter at every homogeneous gap level at a price of $\ln(I{+}1) = O(\ln\ln(W_I/W_0))$ inside the logarithms. What remains open: segments whose stepwise-maximal gap far exceeds their detection level ($B\bar{g}_m \gg G_m$, i.e. strongly gap-*heterogeneous* segments; a very bad expert served while the closest competitor pins the detection level) retain only the instance-adaptive bound; the lower-bound construction of Proposition G.4 is gap-homogeneous and hence says nothing there, and we do not know which side moves. The most promising route we found, and where it breaks: a *moving-candidate* gate; one whose comparator is re-nominated while the confirmation e-process keeps running, made two-sidedly $\mathcal{F}_{t-1}$-measurable so that its adoption-validity certificate still holds (that half *is* provable); combined with successive refinement across the dyadic scales would charge each heterogeneous segment at its own gap; but the power argument then needs the near-optimal cluster to be separated (which re-imports homogeneity through the back door), and adoptions confirmed at one scale perturb the loss prefixes that other scales are betting on, a cross-scale term we cannot bound without assumptions. Both blockers, it turns out, are killed by a protocol change plus one stability condition rather than by cluster separation, and Theorem G.7 below closes the heterogeneous case on that class; fully unstable gap profiles remain open. Proof below; measured in E9 (Appendix H.12).

The protocol change is *candidate locking* (hysteresis). The *successive-refinement* GCL (SR-GCL) is MS-GCL with one modification: scale i refreshes its nominated candidate only at the boundaries of its own W_i-step matured windows; between boundaries the candidate is locked, so nomination oscillation within a near-tied cluster can no longer reset a running confirmation (the first blocker disappears by design, not by assumption). Everything else (candidate-indexed incumbent-varying gates, per-scale budget $\alpha_{\mathrm{meta}}/(I{+}1)$, segment-counter conventions) is unchanged. The added condition replaces cross-expert homogeneity by per-expert stability: call segment m (comparator k_m) *c-level-stable*, for a constant $c \geq 1$ (forced by the two-sided sandwich; the proof uses $c \geq 1$ explicitly), if every expert $j \neq k_m$ has a level $g_j > 0$ with $g_j \leq \mathbb{E}[\ell_t^j - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \leq c\, g_j$ throughout the segment. The levels may be arbitrarily heterogeneous across experts (precisely the regime Theorem G.6 leaves open) but each expert must hold its own level up to the factor c. Write $G_m = \min_{j \neq k_m} g_j$ for the detection level, and require G_m detectable at some grid scale i_m (nomination margin and (†) at scale i_m, as in Theorem G.6), with (†) holding at every coarser grid scale at its own resolution.

Theorem G.7 (Successive refinement: heterogeneous closure on the level-stable class)**.** *Under Assumption G.1, bounded losses, and A2′, SR-GCL satisfies:* ***(a)*** *the adoption-validity certificate of Theorem G.6(a) at total level α_{meta} (locking only reduces candidate resets; the null supermartingale argument is unchanged).* ***(b)*** *For every $\delta, \delta' \in (0,1)$ and every comparator σ with S switches all of whose segments are c-level-stable with detection level G_m detectable at a grid scale, with probability at least $1 - \alpha_{\mathrm{meta}} - \beta_{\mathrm{det}} - \delta - \delta'$, the per-switch excess of segment m is*

$$O\Big(\tfrac{c^3\, B^2 \ln(2K(I+1)T(T+1)/\delta\alpha_{\mathrm{meta}})}{G_m} \;+\; B\,(W_0 + H + 1)\Big),$$

independently of the served gap $\bar{g}_m$: a segment that starts on an arbitrarily bad incumbent ($B\bar{g}_m \gg G_m$) descends through the dyadic scales (each coarse scale evicts incumbents above its own resolution within one locked window) and pays a geometric sum dominated by the finest scale. On the c-level-stable detectable class this matches the certification lower bound of Proposition G.4 at every detection level $G_m \in [\Delta, B]$ up to the factor c^3 and logarithmic terms.

What Theorem G.7 does and does not close. It removes the last worst-case gap of Proposition G.4 on segments whose experts hold stable loss levels; the "very bad expert served while the closest competitor pins the detection level" scenario that Theorem G.6 left open is now priced at $O(B^2/G_m)$, not $O(\bar{g}_m B^3/G_m^2)$. The price is the locking latency (the BW_0 term, with $W_0 = O(\ln)$ at the coarsest useful resolution) and the stability condition itself: an expert whose conditional gap *drifts across dyadic levels within a single comparator segment* violates c-level-stability for small c, and for such unstable profiles only the instance-adaptive bound of Corollary G.5 survives. The lower-bound side matches: Proposition G.4's construction is level-stable (i.i.d.), so it certifies optimality *within* the class; whether unstable profiles are strictly harder is open; we do not know which side moves. Proof below; measured in E11 (Appendix H.12), on a designed stream.

Three boundaries remain. First, Theorem G.3(b) holds on the (Δ, W)-detectable class only: if no expert dominates its segment by a detectable margin, the nomination can oscillate and the gate (which resets on candidate changes) may never accumulate evidence; below the detection threshold the meta-policy holds *only* its certificate (a), not tracking. This is the meta-level replay of the no-free-lunch boundary (Theorem 3.4), and Proposition G.4 shows it is intrinsic to certification, not to GCL. Second, GCL does not subsume Theorem G.2: the Hedge guarantee holds against the best *fixed* expert with no margin condition at all, while GCL's tracking needs the detectable margin; the two theorems trade comparator strength against assumption strength, and the endpoint $\alpha_{\mathrm{meta}} \to 1$ (gate fires instantly) collapses GCL to ungated nomination-following with no certificate. Third, claim (i) of Theorem G.2 (all base ledgers hold in silico) carries over unchanged: the ledgers of *non-incumbent* policies remain per-policy in-silico statements. What *can* be closed is the ledger of the trajectory actually served, and Theorem G.8 closes it. Proofs below; measured in E7 (Appendix H.12).

Call a structural intervention *materialized* if it occurs on the served trajectory: a base-gate firing of the current incumbent during its own tenure, or an adopted meta switch. The key observation is a *tenure-transfer* property of Assumption G.1: the stream is exogenous, so each policy's in-silico pipeline (outputs, recursion state, gate wealth) is the same measurable function of the shared history whether or not that policy is being served; during incumbent k's tenure, the served pipeline *is* policy k's in-silico pipeline, path by path. Base certificates therefore materialize on the served trajectory with no new supermartingale argument, and the two budgets compose:

Theorem G.8 (Composite ledger for the served trajectory)**.** *Under Assumption G.1, bounded losses, and A2′, let each of the K base policies run its own Axiom 2.5 gate family with per-policy budget α_{S}/K (spending within each policy as in Axiom 2.5), and let GCL run with budget α_{meta}. Then*

$$\mathbb{P}\big(\text{some materialized intervention is false}\big) \;\le\; \alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}},$$

uniformly over stopping times, where a materialized base firing is false if its H_0 holds throughout *the firing segment (inheriting the momentary-null exclusion of Theorem 3.1(b): a firing at an instant where the null holds only after being violated earlier in the same segment is not constrained), and a materialized adoption is false in the sense of Theorem G.3(a). Under the alternative convention that keeps every base policy at its deployment budget α_{S} (as in E1–E7), the same statement holds at level $K\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}}$.*

The composite ledger removes the boundary stated after Theorem G.2 for the served trajectory: every structural intervention the deployment actually executes (base or meta) now sits under one joint anytime-valid budget $(\alpha_{\mathrm{S}}, \alpha_{\mathrm{meta}})$. The price is quantified: rebudgeting to α_{S}/K shifts every base gate's detectability frontier (†) by an additive $\ln K$ inside the logarithm (power is paid for validity) and the in-silico ledgers of policies while *not* incumbent remain simulation statements, as before. The composite statement holds with MS-GCL in place of GCL (its Part (a) has the same form at the same total level α_{meta}).

All of Theorems G.2, G.3, G.6 and G.8 lean on Assumption G.1, and it is worth stating precisely what fails without it and what survives a quantified weakening. Two distinct roles must be separated. *Computability* of the in-silico trajectories survives performativity outright: every policy's pipeline is a measurable function of the *realized* history, so on the served run all K pipelines, gates, and ledgers can still be computed, and every supermartingale/Ville argument (which conditions only on the realized filtration) still holds; the in-silico certificates are valid as statements about the realized run no matter how the stream reacts to serving. What Assumption G.1 buys is *counterfactual equality*: that policy k's in-silico trajectory coincides with what would have been realized had k been served. That is the reading under which regret over the library and tenure transfer have deployment meaning, and it admits a quantified relaxation:

Assumption G.9 (ε-performativity)**.** There is a sequence $(\varepsilon_t)_{t\ge 1}$ such that for any two serving rules, the conditional laws of (x_t, y_t) given any common realized history $\mathcal{F}_{t-1}$ differ in total variation by at most ε_t. Write $E_T = \sum_{t\le T} \varepsilon_t$. Assumption G.1 is the case $\varepsilon_t \equiv 0$.

Proposition G.10 (Performativity robustness of the counterfactual readings)**.** *Under Assumption G.9 (in place of Assumption G.1), bounded losses, and A2′: (i) for any two serving rules the induced path laws on horizon T differ in total variation by at most E_T; consequently (ii) every probability-level guarantee of Theorems G.2, G.3, G.6 and G.8, read against the* counterfactual deployment *of the referenced policies (regret against the loss expert k would have realized if served; falseness of a materialized intervention judged in the counterfactual deployment of its policy), holds with its level inflated by $k_{\mathrm{ref}}E_T$, where $k_{\mathrm{ref}} \le K$ is the number of counterfactual deployments the guarantee references ($k_{\mathrm{ref}} = 1$ for a single adopted switch or a single materialized intervention; $k_{\mathrm{ref}} =$ the number of* distinct *policies materializing an intervention on the trajectory for the composite ledger; at most K, and 1 for a run with a single incumbent; $k_{\mathrm{ref}} = K$ for regret over the whole library: $\delta \mapsto \delta + k_{\mathrm{ref}}E_T$, $\alpha_{\mathrm{meta}} \mapsto \alpha_{\mathrm{meta}} + k_{\mathrm{ref}}E_T$, $\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}} \mapsto \alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}} + k_{\mathrm{ref}}E_T$), uniformly over stopping times $\le T$; and (iii) the deterministic per-trajectory ledgers (Theorem 3.1(a); Theorem 3.3; Theorem F.10) hold on the realized run unchanged. The boundary is sharp in kind: under persistent performativity $\varepsilon_t \ge \varepsilon > 0$, E_T grows linearly and the counterfactual readings are vacuous at $T \gtrsim 1/\varepsilon$; the certificates degrade continuously in the total performative drift, not abruptly, but they do die.*

Proposition G.10 is a quantified robustness statement, not a removal of the assumption: it prices how much feedback coupling the meta and composite layers tolerate (E_T must be small next to the budgets) and where they fail. One further boundary: E_T is *not* nonparametrically estimable from observed data; total-variation distance can hide on sets the realized trajectory never visits, so no finite-sample certificate of $\varepsilon_t \le c$ exists without assumptions. What a served/shadow A/B deployment *can* support is a one-sided falsification audit: an e-process on the served-vs-shadow loss differential (the machinery of A3) can *reject* a claimed budget $E_T \le c$ at anytime validity, but never certify it; that falsification audit is the full extent of empirical access to Assumption G.9. Proof below.

Proof of Theorem G.2. Step 0 (simulability and booking). Under Assumption G.1 each base policy's trajectory is a deterministic function of the shared stream history and its own internal randomness (seeded), so the loss vectors $(\ell_t^1, \dots, \ell_t^K)$ are well-defined and observable once y_t matures, regardless of which policy was served; A2′ books each of the $K{+}1$ ledgers exactly once per step, so claim (i) is immediate; each base theorem is a statement about its own booked trajectory, which is unchanged by the meta-policy's existence.

Step 1 (per-instance anytime Hedge). Fix an interleaved instance $i \in \{0, \dots, H\}$; it acts on steps $t \equiv i \pmod{H{+}1}$. Consecutive own steps are $H{+}1 > H$ apart, so at each own step the losses of all previous own steps have matured; the instance is a standard full-information forecaster. Exponential weights with time-varying rate $\eta_n = \sqrt{8 \ln K / n}$ over losses in $[0, B]$ satisfies, deterministically for every number n_i of own steps (Cesa-Bianchi & Lugosi, 2006, Thm. 2.3, rescaled by B), $\sum_{\text{own } t} \bar{\ell}_t - \min_k \sum_{\text{own } t} \ell_t^k \le B\big(\sqrt{2 n_i \ln K} + \sqrt{\ln K / 8}\big)$ (the additive $B\sqrt{\ln K/8}$ is the theorem's constant term, one per instance), where $\bar{\ell}_t = \sum_k p_t(k)\, \ell_t^k$ is the weighted-average loss under the instance's weights p_t.

Step 2 (across instances). Fix any comparator k; summing the Step-1 bounds and using $\min_k \sum_i (\cdot) \ge \sum_i \min_k (\cdot)$ in the correct direction (each instance's bound holds in particular for the common k) gives, for every T with $\sum_i n_i = T$, $\sum_{t \le T} \bar{\ell}_t - \sum_{t \le T} \ell_t^k \le B\sqrt{2 \ln K} \sum_i \sqrt{n_i} + (H{+}1) B \sqrt{\ln K / 8} \le B\sqrt{2(H{+}1) T \ln K} + (H{+}1) B \sqrt{\ln K / 8}$, the last step by Cauchy–Schwarz on the leading sum. Taking the minimum over k yields the second and third terms of the theorem; the additive term is $O(1/T)$ after normalization and is dominated by the sampling term.

Step 3 (sampling deviation, uniformly in T). The served loss is $\ell_t^{\text{meta}} = \ell_t^{k_t}$ with $k_t \sim p_t$, so $M_T = \sum_{t \le T} (\ell_t^{\text{meta}} - \bar{\ell}_t)$ is a martingale with increments in $[-B, B]$; by Hoeffding's lemma, $\exp(\lambda M_T - \lambda^2 T B^2/2)$ is a supermartingale for every fixed $\lambda > 0$. Spend $\delta_j = \delta/(j{+}1)(j{+}2)$ on epoch $j \ge 1$, split $\delta_j/2$ per tail (so the total spend is $\sum_{j \ge 1} \delta_j = \delta/2 \le \delta$): for each tail, Ville's inequality applied to the supermartingale with $\lambda_j = \sqrt{2 \ln(2/\delta_j)/(2^j B^2)}$ gives $\mathbb{P}(\exists T \le 2^j : \pm M_T \ge \ln(2/\delta_j)/\lambda_j + \lambda_j T B^2/2) \le \delta_j/2$, and for $T \le 2^j$ the threshold is at most $B\sqrt{2 \cdot 2^j \ln(2/\delta_j)}$. A given T lies in epoch $j = \lceil \log_2 T \rceil \vee 1 \le j_T$ with $2^j \le 2T$, so on the complement of the union (probability $\ge 1 - \delta$), simultaneously for all T, $|M_T| \le 2B\sqrt{T \ln\big(2(j_T{+}1)(j_T{+}2)/\delta\big)}$ (using $\ln(2/\delta_j) = \ln(2(j{+}1)(j{+}2)/\delta)$, monotone in j): the last term of the theorem after dividing by T. Combining Steps 2 and 3 and dividing by T proves the regret display.

Step 4 (claim (ii)). If base policy k^* is budget-compliant; Theorem 4.2 applied to its in-silico trajectory (valid by claim (i)) bounds $T^{-1} \sum_t \ell_t^{k^*}$ by its average certified risk plus $B_{\text{dec}}[\alpha + O(1/\gamma T)]$; adding the regret display bounds the served loss by the same quantity plus $O(\sqrt{(H{+}1)\ln(K/\delta)/T})$.

The regret is against the best fixed k: no claim is made about switching comparators, and the meta-level switches are not e-gated; Theorem 3.1(b) is inherited only inside each base policy, as stated in the text. □

Proof of Theorem G.3. Losses are normalized by B throughout ($\ell \in [0, 1]$); the displayed bound restores the B factors.

Part (a). On each monitoring segment the gate is the mixture-betting e-process of Axiom 2.5 run on the matured differential $d_t = \ell_t^{\text{inc}} - \ell_t^{\text{cand}} \in [-1, 1]$, which under Assumption G.1 + A2′ is a measurable function of the shared history (both trajectories are in silico). Under the null $\mathbb{E}[d_t \mid \mathcal{F}_{t-1}] \le 0$ every betting factor $1 + \lambda d_t$ ($0 < \lambda \le 1/2$) has conditional mean ≤ 1, so the mixture wealth is a nonnegative supermartingale; Ville's inequality at threshold $j(j{+}1)/\alpha_{\text{meta}}$ spends $\alpha_{\text{meta}}/j(j{+}1)$ on segment j, and the segment index advances on every firing, window expiry, *and candidate change*, so the total spend over all segments is $\sum_j \alpha_{\text{meta}}/j(j{+}1) \le \alpha_{\text{meta}}$ regardless of how often the nomination changes. This is Theorem 3.1(b) with the gate family indexed by (incumbent, candidate) pairs.

Part (b). Fix T, δ, and the comparator σ with segments $m = 1, \dots, S{+}1$, expert k_m on segment m, all (Δ, W)-detectable at level δ by horizon T. Work on the intersection of three events: E_1 = no false adoption (Part (a), probability $\ge 1 - \alpha_{\text{meta}}$); E_2 = every gate monitoring a differential with conditional mean $\ge \Delta$ satisfying (†) fires within W steps of its monitoring-segment start

(the power condition of Theorem C.1 transcribed to the meta gate; total spending β_{det}); E_3 = uniform concentration of all length-$\le W$ trailing-window loss sums: for each expert k and each window end time $t \le T$, allocate $\delta/(K\,t(t{+}1))$; by Azuma–Hoeffding with increments in $[0,1]$, on E_3 every trailing window sum of length $n \le W$ deviates from its cumulative conditional mean by at most $\sqrt{(n/2)\ln\big(2Kt(t{+}1)/\delta\big)} \le \sqrt{(W/2)\ln\big(2KT(T{+}1)/\delta\big)}$, and the total spend is $\le \delta$. $\mathbb{P}(E_1 \cap E_2 \cap E_3) \ge 1 - \alpha_{\mathrm{meta}} - \beta_{\mathrm{det}} - \delta$.

On this event, decompose each comparator segment m into three phases and bound the per-step excess $\ell_t^{\mathrm{GCL}} - \ell_t^{k_m} \le 1$ during the first two.

Phase 1 (nomination). Consider any step t whose trailing nomination window (the last W matured steps) lies entirely inside segment m; this holds once t exceeds the segment start by $W + H + 1$ (the window must both fill with in-segment steps and mature). For any $j' \ne k_m$, the window sums satisfy, on E_3,

$$L_W(j') - L_W(k_m) \;\ge\; \Delta W - 2\sqrt{(W/2)\ln\tfrac{2KT(T+1)}{\delta}} \;>\; 0,$$

where the last inequality is exactly the nomination margin $\Delta \ge \sqrt{2\ln(2KT(T{+}1)/\delta)/W}$ (in normalized units). Hence the windowed leader is k_m at every such step: the nomination is k_m, and stays k_m, from at most $W + H + 1$ steps into the segment. Note the sliding window is reset-free: there are no epochs, so no restart events other than the comparator's own switches.

Phase 2 (confirmation). Once the nomination is k_m and stable, the gate monitors $d_t = \ell_t^{\mathrm{inc}} - \ell_t^{k_m}$ with conditional mean $\ge \Delta$ (the incumbent is some $j' \ne k_m$); by E_2 it fires within W steps plus at most $H{+}1$ steps of maturation lag, and the adopted switch installs k_m. (If the incumbent is already k_m at the segment start, Phases 1–2 cost nothing.)

Phase 3 (installed). While the incumbent is k_m the per-step excess over the comparator is zero, and on $E_1 \cap E_3$ it is never removed within the segment: by Phase 1 no other expert is nominated once the window is inside the segment, and any monitoring segment opened against a candidate $j' \ne k_m$ during the boundary steps has differential with conditional mean $\le -\Delta < 0$, so its adoption is blocked on E_1.

Each comparator segment therefore incurs at most $(W + H + 1) + (W + H + 1) = 2W + 2H + 2$ steps of unit (normalized) excess, and there are $S{+}1$ segments; multiplying by B and dividing by T gives the display. □

Proof of Proposition G.4. Normalize $B = 1$. Two experts with i.i.d. losses: under the environment $\mathcal{E}$, expert 1 has Bernoulli$(1/2)$ losses and expert 2 Bernoulli$(1/2 - \Delta)$ losses, independent across steps; the meta-policy starts on expert 1 (the environment can always relabel so that the incumbent is the worse expert: the certification constraint is asymmetric (adoptions must be certified while retentions are free) so the policy cannot hedge against the relabeling by switching without evidence, and placing the gap on the non-incumbent is without loss of generality). Consider the null environment $\mathcal{E}_0$ in which both experts are Bernoulli$(1/2)$. An α_{meta}-certified meta-policy must satisfy $\mathbb{P}_{\mathcal{E}_0}(\text{adopt expert 2 ever}) \le \alpha_{\mathrm{meta}}$, since under $\mathcal{E}_0$ the differential has conditional mean $0 \le 0$ (the null of Theorem G.3(a) holds forever).

Let A_n = {adopt by step n}, a measurable event of the first n differentials. The likelihood ratio of $\mathcal{E}$ to $\mathcal{E}_0$ on n steps of expert-2 losses satisfies $\mathbb{E}_{\mathcal{E}}[\log \mathrm{LR}_n] = n\,\mathrm{KL}\big(\mathrm{Ber}(1/2 - \Delta) \,\|\, \mathrm{Ber}(1/2)\big) \le 4n\Delta^2$ for $\Delta \le 1/4$, with increments bounded by $\log\frac{1/2+\Delta}{1/2-\Delta} \le 8\Delta$. By Azuma–Hoeffding, $\mathbb{P}_{\mathcal{E}}\big(\log \mathrm{LR}_n > 4n\Delta^2 + 8\Delta\sqrt{2n\ln(1/\epsilon)}\big) \le \epsilon$, so by change of measure,

$$\mathbb{P}_{\mathcal{E}}(A_n) \;\le\; e^{\,4n\Delta^2 + 8\Delta\sqrt{2n\ln(1/\epsilon)}}\,\mathbb{P}_{\mathcal{E}_0}(A_n) + \epsilon \;\le\; e^{\,4n\Delta^2 + 8\Delta\sqrt{2n\ln(1/\epsilon)}}\,\alpha_{\mathrm{meta}} + \epsilon.$$

Take $\epsilon = 1/4$ and $n^* = c_0 \ln(1/2\alpha_{\mathrm{meta}})/\Delta^2$ with c_0 a small universal constant chosen so that, for $\alpha_{\mathrm{meta}} \le 1/8$, the exponent is at most $\frac{1}{2}\ln(1/2\alpha_{\mathrm{meta}})$; then $\mathbb{P}_{\mathcal{E}}(A_{n^*}) \le e^{\frac{1}{2}\ln(1/2\alpha_{\mathrm{meta}})}\alpha_{\mathrm{meta}} + \frac{1}{4} = \sqrt{\alpha_{\mathrm{meta}}/2} + \frac{1}{4} \le \frac{1}{2}$ for $\alpha_{\mathrm{meta}} \le 1/8$. Hence with probability $\ge 1/2$ under $\mathcal{E}$ the meta-policy serves expert 1 for the first $\min\{T, n^*\}$ steps, paying expected excess Δ per step: expected excess $\ge \frac{1}{2}\Delta\min\{T, c_0\ln(1/2\alpha_{\mathrm{meta}})/\Delta^2\}$, which is the claim with $c = c_0/2$. The construction uses only the certification property, not any detail of GCL. □

Proof of Corollary G.5. Work in normalized units ($\ell \in [0,1]$) on the intersection of $E_1 \cap E_2 \cap E_3$ from the proof of Theorem G.3(b) and a fourth event E_4 below. On $E_1 \cap E_2 \cap E_3$, the phase analysis shows that within each comparator segment m the excess-incurring steps form an *initial stretch* of the segment of some random length $n_m \le 2W + 2H + 2$ (Phases 1–2; once k_m is installed it is never removed within the segment). During those steps the incumbent is some expert $j' \neq k_m$ and the per-step excess is $d_t = \ell_t^{j'(t)} - \ell_t^{k_m}$ with increments in $[-1, 1]$. Anchor partial sums at the segment start: for each segment m and each prefix length $n \le 2W + 2H + 2$, Azuma–Hoeffding with spending $\delta'/\big((S{+}1)(2W{+}2H{+}2)\big)$ gives

$$\sum_{t \text{ in first } n \text{ steps}} d_t \;\le\; \sum_t \mathbb{E}[d_t \mid \mathcal{F}_{t-1}] + \sqrt{2n \ln \tfrac{2(S+1)(2W+2H+2)}{\delta'}}$$

$$\le\; \bar{g}_m\, n + \sqrt{2(2W{+}2H{+}2) \ln \tfrac{2(S+1)(2W+2H+2)}{\delta'}},$$

using the definition of $\bar{g}_m$: each conditional mean satisfies $\mathbb{E}[d_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\ell_t^{j'(t)} - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \le \max_{j' \neq k_m} \mathbb{E}[\ell_t^{j'} - \ell_t^{k_m} \mid \mathcal{F}_{t-1}]$ pointwise ($j'(t)$ is $\mathcal{F}_{t-1}$-measurable, so it may be replaced by the stepwise maximum term by term), and the sum of the stepwise maxima over the first n steps is at most $\bar{g}_m n$ by definition. The incumbent identity $j'(t)$ may change during the stretch (an intermediate adoption of a third expert is possible and need not be false); because the maximum sits *inside* the sum, the bound is indifferent to which $j' \neq k_m$ is served at each step. Let E_4 be the intersection of these events; the total spend is $\le \delta'$, so $\mathbb{P}(E_1 \cap E_2 \cap E_3 \cap E_4) \ge 1 - \alpha_{\text{meta}} - \beta_{\text{det}} - \delta - \delta'$. Summing over the $S{+}1$ segments and restoring B gives the display. The trivial bound $d_t \le 1$ recovers Theorem G.3(b), so the corollary is never worse than the theorem beyond the deviation term. □

Proof of Theorem G.6. Normalized units ($\ell \in [0,1]$) throughout; the display restores B.

Part (a). Each scale's gate monitors $d_t = \ell_t^{\text{inc}_t} - \ell_t^{\text{cand}}$. The incumbent index inc_t is determined by past firings and nominations, hence $\mathcal{F}_{t-1}$-measurable, and both trajectories are in silico under Assumption G.1 + A2′; under the null $\mathbb{E}[d_t \mid \mathcal{F}_{t-1}] \le 0$ every betting factor has conditional mean ≤ 1, so the mixture wealth is a nonnegative supermartingale *even though the incumbent may change while the segment is open*. Ville at threshold $j(j{+}1)(I{+}1)/\alpha_{\text{meta}}$ spends $\alpha_{\text{meta}}/\big((I{+}1)j(j{+}1)\big)$ on scale i's segment j; the segment counter advances on candidate change, firing, and window expiry, so the spend per scale is $\le \alpha_{\text{meta}}/(I{+}1)$ and the total over the $I{+}1$ scales is $\le \alpha_{\text{meta}}$.

Part (b). Fix T, δ, δ', and the comparator. Work on $E_1 \cap E_2 \cap E_3 \cap E_4$: E_1 = no false adoption at any scale (Part (a), $\ge 1 - \alpha_{\text{meta}}$); E_2 = every gate monitoring a differential with conditional mean $\ge$ its scale's detectability level and satisfying (†) at its scale fires within its window (total spend β_{det}); E_3 = uniform Azuma concentration of all trailing-window sums at all scales, with allocation $\delta/\big(K(I{+}1)\,t(t{+}1)\big)$ per (expert, scale, window end); deviations at scale i at most $\sqrt{(W_i/2)\ln(2K(I{+}1)T(T{+}1)/\delta)}$, total spend $\le \delta$; E_4 = the prefix-anchored Azuma events of Corollary G.5 at stretch cap $2W_{i_m} + 2H + 2$ per segment (spend $\le \delta'$; the $\ln$ uses the largest cap $2W_I + 2H + 2$).

Fix a comparator segment m with detection level G_m at grid scale i_m. *Nomination at scale i_m.* Once the trailing W_{i_m} matured window lies inside the segment ($W_{i_m} + H + 1$ steps in), E_3 and the margin condition $G_m \ge \sqrt{2\ln(2K(I{+}1)T(T{+}1)/\delta)/W_{i_m}}$ make k_m the scale-i_m leader at every subsequent step of the segment, exactly as in Phase 1 of Theorem G.3(b). *Confirmation at scale i_m.* From that step on, scale i_m's candidate is stably k_m (or the incumbent already is k_m, costing nothing), so its candidate-indexed gate runs without candidate resets; its differential $\mathbb{E}[\ell_t^{\text{inc}_t} - \ell_t^{k_m} \mid \mathcal{F}_{t-1}] \ge G_m$ *whatever* the current incumbent $\neq k_m$ is (the detectability level is uniform over $j \neq k_m$), and (the point of candidate indexing) an intermediate adoption at another scale does not reset it. By E_2 it fires within W_{i_m} steps plus $H{+}1$ of maturation, installing k_m. One edge case: scale i_m's gate segment may already be open at stabilization with candidate k_m (nominated before the margin stabilized), so no candidate-change reset marks the start of the clean stretch; but window expiry advances the segment counter, so the straddling segment expires within W_{i_m} steps of stabilization and the next segment runs a full window entirely under the $\ge G_m$ differential and fires; the total is still within the $2W_{i_m} + 2H + 2$ stretch cap. (†) is invoked here for a differential whose incumbent side

varies; as noted after the theorem, the wealth-growth argument of Theorem C.1 uses only the uniform conditional-mean lower bound over $j \neq k_m$, so the varying composition is immaterial. *Intermediate adoptions.* While the stretch is running, other scales may adopt other experts; each such adoption changes inc_t but neither resets scale i_m's gate nor extends the stretch bound, and the per-step excess against k_m is controlled by the stepwise-maximal machinery of Corollary G.5 regardless of which $j' \neq k_m$ is served: on E_4, the excess over the stretch of length $n_m \leq 2W_{i_m} + 2H + 2$ is at most $\bar{g}_m n_m$ plus the deviation term. *Stability.* Once k_m is installed, every candidate gate at every scale monitors a differential with conditional mean $\leq -G_m < 0$, so on E_1 no adoption removes k_m within the segment; wrong nominations at scales too small for the margin are likewise blocked on E_1. Summing over segments and restoring B gives the display. For the homogeneity claim: c-gap-homogeneity bounds every stepwise maximum by cG_m/B (normalized), hence $\bar{g}_m \leq cG_m/B$, and $W_{i_m} \leq 2 \cdot 2B^2 \ln(2K(I{+}1)T(T{+}1)/\delta)/G_m^2$ by dyadicity of the grid (the smallest admissible scale at most doubles the minimal admissible window), assuming (†) at that scale holds at the same order; as at the base gates, (†)'s sample requirement is $\Theta(B^2 \ln((I{+}1)/\alpha_{\mathrm{meta}})/G_m^2)$. The per-switch cost is then $cG_m \cdot O(B^2 \ln(KIT/\delta\alpha_{\mathrm{meta}})/G_m^2) = O(cB^2 \ln(KIT/\delta\alpha_{\mathrm{meta}})/G_m)$, maximized over the class at $G_m = \Delta$. □

Proof of Theorem G.7. Normalized units ($\ell \in [0,1]$, levels $g_j \in [0,1]$, resolutions $\varepsilon_i := \sqrt{2\ln(2K(I{+}1)T(T{+}1)/\delta)/W_i}$) throughout; the display restores B. Note $\varepsilon_i = \varepsilon_{i+1}\sqrt{2}$ and $\varepsilon_i W_i = 2\ln(2K(I{+}1)T(T{+}1)/\delta)/\varepsilon_i$.

Part (a). Identical to Theorem G.6(a): each scale's gate monitors the incumbent-varying differential against an $\mathcal{F}_{t-1}$-measurable locked candidate; under the null the mixture wealth is a nonnegative supermartingale, and locking only changes *when* the candidate index can change (at window boundaries instead of every step), which affects the segment counter's advancement times but not the Ville threshold accounting: the spend per scale remains $\leq \alpha_{\mathrm{meta}}/(I{+}1)$, total $\leq \alpha_{\mathrm{meta}}$.

Part (b). Work on $E_1 \cap E_2 \cap E_3 \cap E_4$ as in Theorem G.6, with E_3 the uniform Azuma concentration of all trailing-window sums at all scales and all window boundaries (allocation $\delta/(K(I{+}1)t(t{+}1))$ per (expert, scale, boundary); deviation at scale i at most $W_i\varepsilon_i/2$ per trailing sum). Fix a c-level-stable segment m.

Step 1 (locked-candidate quality). Let scale i lock candidate $\hat{\jmath}$ at a boundary whose trailing W_i matured window lies inside the segment. On E_3, $\hat{\jmath}$'s realized trailing mean is within $\varepsilon_i/2$ of its conditional-mean average, and likewise for k_m; since $\hat{\jmath}$ is the realized leader, its conditional-mean average gap to k_m is at most ε_i. By c-level-stability the average dominates the level: $g_{\hat{\jmath}} \leq \varepsilon_i$. So every candidate locked by scale i (with an in-segment window) has level at most ε_i; consequently, on E_1-adopted switches, *every adoption during the segment installs an incumbent of level at most the adopting scale's resolution*.

Step 2 (eviction). Suppose during one of scale i's locked windows the incumbent level satisfies $g_{\mathrm{inc}_t} \geq 4c\,\varepsilon_i$ at every step. The monitored differential has conditional mean $\mathbb{E}[\ell_t^{\mathrm{inc}_t} - \ell_t^{\hat{\jmath}} \mid \mathcal{F}_{t-1}] \geq g_{\mathrm{inc}_t} - c\,g_{\hat{\jmath}} \geq 4c\varepsilon_i - c\varepsilon_i \geq \varepsilon_i$ uniformly over whichever incumbent is served (Step 1 bounds the candidate; level-stability bounds both sides; the uniform lower bound is all the wealth-growth argument of (†) uses, as in Theorem G.6). By E_2, (†) at scale i fires within the window: the incumbent is replaced by a level-$\leq \varepsilon_i$ candidate. Contrapositive: after at most two of scale i's locked windows inside the segment (one to align the lock, one to confirm) plus $H{+}1$ maturation steps, some step had $g_{\mathrm{inc}} < 4c\varepsilon_i$. Containment after this drop needs one more accounting, because a gate's wealth may have been accumulated against earlier, worse incumbents. (i) On E_1, any adoption whose monitored segment lies entirely after the drop installs a level below c times some concurrent incumbent level (positive conditional mean at some step forces $g_{\mathrm{cand}} < c\,g_{\mathrm{inc}}$ there by level-stability), hence $\leq c \cdot 4c\varepsilon_i = 4c^2\varepsilon_i$. (ii) A gate whose wealth predates the drop can fire once more (a *stale* firing) installing a level capped only by its scale's resolution (Step 1); but firing resets the gate, and a second firing needs fresh positive-mean evidence on E_1, so each scale contributes at most one stale install per descent. A stale install at level $\leq \varepsilon_{i'}$ is itself evicted, by the coarsest scale k with $4c\varepsilon_k \leq \varepsilon_{i'}$, within $O(W_k) = O(c^2 \ln(2K(I{+}1)T(T{+}1)/\delta)/\varepsilon_{i'}^2)$ steps at per-step excess $\leq c\,\varepsilon_{i'}$: a bill of $O(c^3 \ln(2K(I{+}1)T(T{+}1)/\delta)/\varepsilon_{i'})$ per stale install; with at most one per scale and $\varepsilon_{i'}$ geometric, the total stale bill over the descent is $O(c^3 \ln(2K(I{+}1)T(T{+}1)/\delta)/\varepsilon_{i_m}) = O(c^3 \ln(2K(I{+}1)T(T{+}1)/\delta)/G_m)$, the same order as the main term, and is booked there.

Step 3 (descent). Apply Step 2 at scales $i = 0, 1, \ldots, i_m$ in order. Phase i (from the first boundary after phase $i-1$ ends until scale i's eviction criterion is met) lasts at most $2W_i + H + 1$ steps, during which the incumbent level is at most $4c^2\varepsilon_{i-1} = 4\sqrt{2}\,c^2\varepsilon_i$ by Step 2(i) applied to phase $i-1$, stale installs being billed separately by Step 2(ii) (for $i = 0$: at most 1). The conditional-mean excess per step is at most $c\,g_{\text{inc}} \le 4\sqrt{2}\,c^3\varepsilon_i$; summing, phase i costs $O\big(c^3\varepsilon_i W_i\big) = O\big(c^3 \ln(2K(I{+}1)T(T{+}1)/\delta)/\varepsilon_i\big)$, and the initial phase costs $O(W_0 + H + 1)$. Since ε_i shrinks geometrically, $\sum_{i \le i_m} 1/\varepsilon_i = O(1/\varepsilon_{i_m}) = O(1/G_m)$ (dyadicity: $\varepsilon_{i_m} \ge G_m/\sqrt{2}$ at the smallest admissible scale).

Step 4 (terminal phase and stability). At scale i_m, $\varepsilon_{i_m} \le G_m \le g_j$ for every $j \ne k_m$, so by Step 1 the only candidate scale i_m can lock is k_m itself (if some g_j sits exactly at the resolution, pass to the next finer grid scale, which is admissible by dyadicity at the cost of the constant already absorbed); its gate then evicts any incumbent of level $\ge 4c\varepsilon_{i_m}$ and installs k_m. Once k_m is incumbent, every gate at every scale monitors a differential with conditional mean $\le -G_m + c\varepsilon_\cdot \cdot \mathbf{1}\{\text{candidate} \ne k_m\}$; candidates other than k_m have level $\ge G_m$, so no differential can sustain a positive mean and on E_1 no adoption removes k_m. Residual excess within the phase before the final install is covered by Step 3's phase-i_m term. The prefix-anchored deviation term (E_4, spend δ') is $O(\sqrt{W_{i_m} \ln((S{+}1)W_I/\delta')})$ per segment, dominated by the main term. Summing the phases and the stale bill of Step 2(ii), restoring B, and absorbing constants gives the per-switch excess $O\big(c^3B^2 \ln(2K(I{+}1)T(T{+}1)/\delta\alpha_{\text{meta}})/G_m + B(W_0 + H + 1)\big)$ of the display. □

Proof of Proposition G.10. (i). Let P and Q be the path laws on horizon T induced by two serving rules. Both decompose over the same filtration; the serving decisions themselves are measurable functions of the realized history (deterministic protocols, or share common exogenous randomization), so the two path laws differ only through the stream kernels. By the chain rule for total variation via step-wise maximal coupling: construct (y_t^P, y_t^Q) on a common space with $\mathbb{P}(y_t^P \ne y_t^Q \mid \text{histories agree}) \le \varepsilon_t$, and note that once the histories agree at $t-1$, the conditional laws at t are within ε_t by Assumption G.9; hence

$$\mathrm{TV}(P, Q) \;\le\; \mathbb{P}(\exists\, t \le T : y_t^P \ne y_t^Q) \;\le\; \sum_{t \le T} \varepsilon_t \;=\; E_T.$$

(ii). Fix a probability-level guarantee, say $\mathbb{P}_{\text{cf}}(A) \le \alpha$ where A is the bad event (a false adoption; regret exceedance) and $\mathbb{P}_{\text{cf}}$ is the law of the counterfactual deployment under which the guarantee's supermartingale argument applies unchanged (under $\mathbb{P}_{\text{cf}}$ the referenced policy is served, so its in-silico and served trajectories coincide by construction and the Assumption G.1-based proof goes through with $\varepsilon \equiv 0$ replaced by conditioning under $\mathbb{P}_{\text{cf}}$; the wealth arguments only use the realized filtration of that run). The event A, expressed as a measurable function of the path, then satisfies under the actual served law $\mathbb{P}_{\text{serve}}(A) \le \mathbb{P}_{\text{cf}}(A) + \mathrm{TV}(\mathbb{P}_{\text{serve}}, \mathbb{P}_{\text{cf}}) \le \alpha + E_T$ by (i). Uniformity over stopping times $\le T$ is inherited because the bad events are already stopping-time-uniform under $\mathbb{P}_{\text{cf}}$ and the TV bound is at the path level. For guarantees referencing k_{ref} counterfactual deployments (regret over the library, $k_{\text{ref}} = K$; the composite ledger, one per *distinct* policy materializing an intervention: couplings are per policy, not per intervention, so repeated interventions by the same incumbent cost one coupling), couple the served run simultaneously with each referenced counterfactual run pairwise as in (i): on the intersection of the coupling events (probability at least $1 - k_{\text{ref}}E_T$ by a union bound) every referenced counterfactual path coincides with its Assumption G.1-proof counterpart, so the guarantee's bad event transfers at total inflation $k_{\text{ref}}E_T$; we do not claim the single-E_T rate for multi-reference guarantees. *(iii).* The deterministic ledgers are identities of the realized run: they hold path by path and reference no counterfactual, so Assumption G.9 plays no role. □

Proof of Theorem G.8. Two ingredients, both already proved. *(Tenure transfer.)* Under Assumption G.1 the outcome stream (y_t) is exogenous: it does not depend on which policy is served. By A2′, policy k's pipeline (forecasts, booked quantiles, recursion state, gate wealth processes) is a deterministic measurable function of the shared history $(y_s)_{s<t}$ and policy k's own protocol, hence identical in silico and in service, path by path. Therefore every materialized base firing (a firing of incumbent k's gate during k's tenure) *is* a firing of policy k's in-silico gate process, and the event "some materialized base firing is false" is contained in the event "some in-silico base firing of some policy is false".

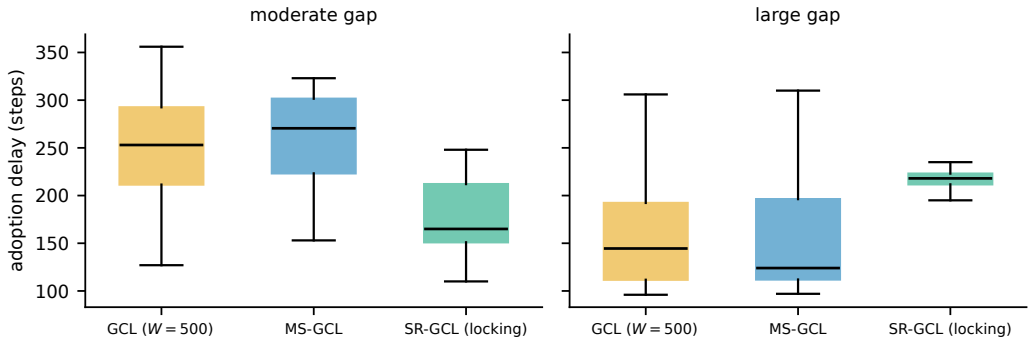

Figure 4: **Meta-layer ablation: post-onset adoption delay by gap level** (E9/E11, linear backbone; boxes = interquartile range over runs with a post-onset adoption, whiskers = range without fliers; each run contributes its first adopted switch). At a moderate gap (1.5σ, near-tied post-onset cluster, left) SR-GCL's candidate locking confirms fastest: oscillating nominations can no longer reset a running confirmation; at a large gap (6σ, one dominant expert, right) locking *hurts*: a locked coarse scale waits for its own window boundary to refresh a stale candidate, the locking latency that Theorem G.7's $O(BW_0)$-per-phase terms price. Both sides of the trade appear in the runs.

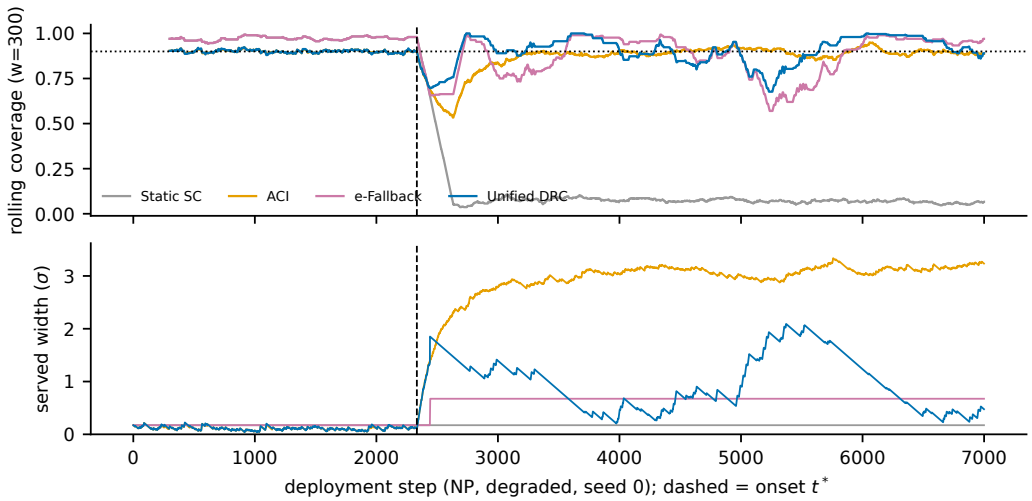

Figure 5: Case study (NP, degraded, seed 0): rolling coverage (top, window 300) and served width (bottom). At onset (dashed), static conformal collapses; ACI recovers coverage by inflating its width many-fold relative to its own pre-onset width; the unified policy fires one gated fallback and returns to budget at a fraction of that width.

(Union of Ville budgets.) Apply Theorem 3.1(b) to each policy k with its gate family rebudgeted to total spend α_S/K (the spending schedule within the family is unchanged, only the level is scaled): $\mathbb{P}(\text{some false in-silico firing of policy } k) \le \alpha_S/K$. Union over k gives α_S. Theorem G.3(a) bounds false adoptions by α_{meta}. A final union bound gives $\alpha_S + \alpha_{\text{meta}}$. Every constituent bound is a whole-trajectory statement (Ville's inequality controls the running supremum), so the composite bound holds simultaneously at every stopping time. For the alternative convention, keep each policy at budget α_S and the union over k costs $K\alpha_S$; nothing else changes. □

H Full Experimental Results

This appendix expands the experiments of Section 5 into their protocols, per-stream breakdowns, and significance tests, organized by experiment.

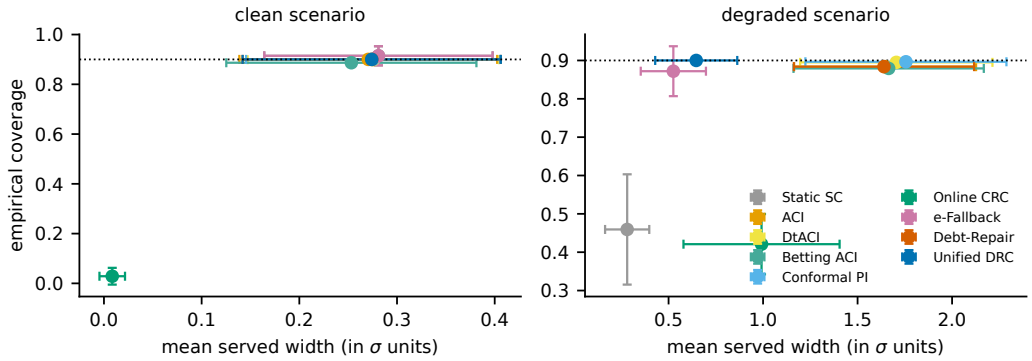

Figure 6: Coverage–width operating points pooled over six streams (10 seeds; error bars = sd across streams). Dotted line = budget $1 - \alpha$. Under degradation both the unified policy and conformal PI sit on the budget line; only the unified policy does so at sub-σ width.

H.1 E1: THE PRICE OF BUDGET COMPLIANCE UNDER DEGRADATION

Compliance is achievable by recalibration alone: at $2.7\times$ the width. Table 1 pools the six streams. We quantify budget compliance by the threshold-free coverage shortfall $(0.90 - \text{cov})_+$ per run, avoiding any ad-hoc violation cutoff. On this axis the two best policies are statistically inseparable in practice: the unified policy (mean 0.001 / max 0.006 over 60 runs) and conformal PI (mean 0.003 / max 0.005); both dominate everything else (DtACI: mean 0.005 / max 0.011; ACI and debt-repair: max 0.022; betting-ACI: max 0.024; fallback-only: max 0.129; the cutoff grid in Table 2 agrees; unified and PI are jointly at zero violations for cutoffs ≤ 0.89, PI is strictly better at 0.895, and both dominate the rest at every cutoff; Figure 7 plots the same axis per scenario). The separation is on the *price* axis: conformal PI holds the budget at width 1.76, DtACI at 1.70, betting-ACI at 1.66 (coverage 0.879, below budget), ACI at 1.65, the unified policy at 0.65; 2.5–$2.7\times$ narrower than every coverage-targeting alternative on all six streams (exact Wilcoxon (Wilcoxon, 1945) $p = 0.031$ at the n=6 floor; Holm-adjusted (Holm, 1979) cluster-bootstrap $p < 10^{-3}$ against each); with lower graded risk (0.016 vs DtACI's 0.024 and PI's 0.023). This is consistent with the coupled-ledger picture of Theorem 3.3: a calibration actuator alone must absorb the entire structural degradation into width, while a certified structural change absorbs it at cost bounded by the intervention ledger. ACI sits at 0.884 ± 0.007 over the horizon: the post-onset transient is the $O(1/\gamma T)$ ledger slack of Theorem 3.1(a); PI's integrator buys back most of that transient (0.897 ± 0.001) with yet more width—the efficiency (not validity) role the framework assigns it. The fallback-only case reaches 0.872 ± 0.065: correct on the streams where the baseline is competitive, but with $\sim 10\times$ ACI's dispersion (0.065 vs 0.0065), because a static-width fallback has no calibration actuator to absorb residual mismatch. The unified policy's advantage over it in coverage ($+0.028$) is *not* significant at the stream level (cluster-bootstrap $p \approx 0.28$); its advantage is the variance collapse and the shortfall profile (max 0.006 vs 0.129). Its coverage edge over PI ($+0.003$) is neither significant at the stream level nor operationally meaningful: we do not lean on it; the frontier claim is the result.

Table 2: Threshold sensitivity: fraction of the 60 degraded runs with coverage below each cutoff. Unified and conformal PI are jointly at zero for cutoffs ≤ 0.89; PI is strictly better at 0.895; both dominate all other policies at every cutoff. No single cutoff carries the comparison.

policy	0.88	0.885	0.89	0.895	0.9
Static SC	1.00	1.00	1.00	1.00	1.00
ACI (R1)	0.35	0.78	0.83	0.83	1.00
DtACI	0.00	0.00	0.02	0.47	0.98
Betting ACI	0.82	1.00	1.00	1.00	1.00
Conformal PI (R2)	0.00	0.00	0.00	0.00	1.00
Online CRC (R3)	1.00	1.00	1.00	1.00	1.00
e-Fallback (R6)	0.33	0.45	0.50	0.50	0.50
Debt-Repair (R4)	0.32	0.78	0.83	0.83	1.00
Unified DRC	0.00	0.00	0.00	0.17	0.33

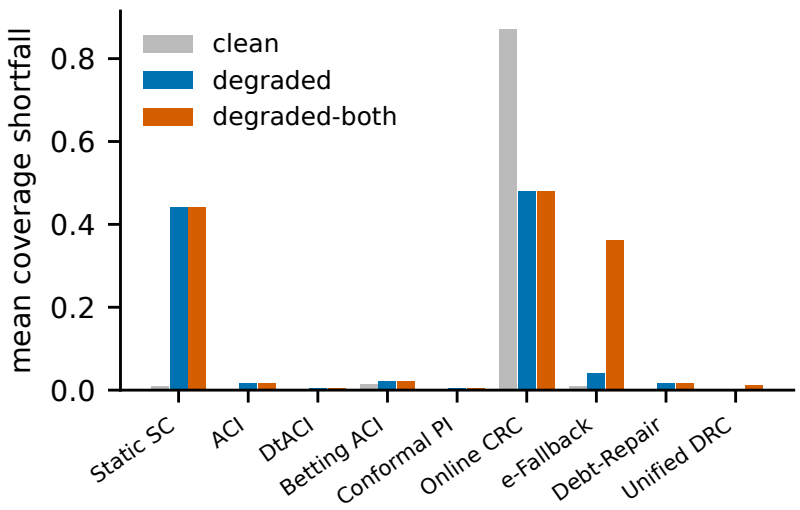

Figure 7: Mean coverage shortfall $(0.90 - \text{cov})_+$ per policy and scenario, over 60 runs each.

Table 3: Degraded-scenario coverage per stream (mean over 10 seeds).

policy	BE	DE	ETTh1	FR	NP	PJM
Static SC	0.777	0.438	0.377	0.379	0.373	0.412
ACI (R1)	0.898	0.881	0.880	0.879	0.879	0.885
DtACI	0.897	0.894	0.896	0.892	0.894	0.896
Betting ACI	0.884	0.880	0.876	0.879	0.878	0.879
Conformal PI (R2)	0.899	0.896	0.896	0.896	0.896	0.896
Online CRC (R3)	0.251	0.491	0.431	0.466	0.453	0.432
e-Fallback (R6)	0.783	0.884	0.938	0.785	0.903	0.939
Debt-Repair (R4)	0.898	0.882	0.880	0.880	0.879	0.885
Unified DRC	0.898	0.902	0.902	0.894	0.901	0.903

H.2 E1: WHAT THE UNIFIED POLICY PAYS ON CLEAN STREAMS

On clean streams the unified policy books 43.5 cost units of repairs (debt-triggered repairs that fire on transient coverage dips), which pure recalibration avoids at the same coverage (0.900) and essentially the same width (0.274 vs ACI's 0.271). The repair operator is not a no-op on clean data: it rolling-median-smooths the clean input view, so each triggered repair delivers a genuinely different forecast; the measured conclusion is that these operations neither help nor hurt on clean data while costing 43.5 units. The e-gate is clean: zero false fallbacks across all 60 clean runs (consistent with Theorem 3.1(b) at joint spend $\leq \alpha_S = 0.05$), so the entire clean-stream overhead is attributable to the ungated debt trigger; the framework says this is the one actuator class for which only deterministic (not anytime-valid) accounting is possible, and the measurement agrees.

H.3 E1: THE RISK FUNCTIONAL DECIDES WHAT IS PROTECTED

The online-CRC special case controls the graded excess-loss risk it recurses on: booked graded risk 0.058 clean and 0.110 degraded against budget 0.10, i.e. within the $O(1/\gamma T)$ ledger slack of Theorem 3.1(a) (the post-onset transient), not a violation; while its *coverage* collapses (0.029 clean): a controller inherits the guarantee of its risk functional and nothing else. This is the empirical face of Proposition D.3: coverage and graded risk are different projections of the same ledger, and choosing the wrong projection forfeits the one you cared about, with no warning.

Table 4: Unified vs. each special case, degraded scenario: mean paired difference over six streams (per-stream seed-means); * = significant at 0.05 after Holm correction on the paired cluster bootstrap (exact Wilcoxon shown for reference; n=6 floor p=0.031).

vs. policy	Δcov.	Δwidth	Δrisk
Static SC	+0.441*	+0.365*	-0.264*
ACI (R1)	+0.016*	-0.999*	-0.014*
DtACI	+0.005*	-1.059*	-0.008*
Betting ACI	+0.021*	-1.018*	-0.012*
Conformal PI (R2)	+0.003	-1.109*	-0.008*
Online CRC (R3)	+0.479*	-0.346	-0.095*
e-Fallback (R6)	+0.028	+0.121*	-0.014*
Debt-Repair (R4)	+0.016*	-0.993*	-0.014*

H.4 E1: REMOVING FALLBACK'S ADVANTAGE

The degraded-both scenario corrupts both input views, so switching to the baseline no longer restores a clean pipeline. Fallback-only collapses (0.539 ± 0.124, shortfall mean 0.361): its one structural actuator is now useless and it has no calibration actuator to compensate. The unified policy partially recovers (0.887 ± 0.005) because its A1 recursion widens on top of whichever pipeline is served, but it does *not* win this scenario: recalibration-only conformal PI attains slightly higher coverage (0.897, higher on all six streams; Holm-adjusted cluster-bootstrap $p < 10^{-3}$) with comparable width, and the unified policy still pays 508 cost units for structural actions that cannot help. When no structural actuator is useful, the unified policy's extra machinery is pure overhead; the deterministic ledger (a) is what protects coverage here, and it belongs to the calibration actuators (Table 5).

Table 5: Degraded-both scenario (both input views corrupted), pooled over six streams. Recalibration-only rows coincide with the degraded scenario by construction (only fallback-touching policies observe the baseline view).

policy	cov.	width	cost	shortfall
Static SC	0.459 ±0.144	0.28	0	0.441/0.535
ACI (R1)	0.884 ±0.007	1.65	0	0.016/0.022
DtACI	0.895 ±0.002	1.70	0	0.005/0.011
Betting ACI	0.879 ±0.003	1.66	0	0.021/0.024
Conformal PI (R2)	0.897 ±0.001	1.76	0	0.003/0.005
Online CRC (R3)	0.421 ±0.079	0.99	0	0.479/0.658
e-Fallback (R6)	0.539 ±0.124	0.41	468	0.361/0.527
Debt-Repair (R4)	0.884 ±0.007	1.64	67	0.016/0.022
Unified DRC	0.887 ±0.005	1.49	508	0.013/0.020

H.5 E1: ABLATION WITHIN THE LEDGER

Table 4 decomposes the degraded-scenario comparison. Against every calibration-only case the unified policy is simultaneously higher-coverage, narrower, and lower-risk (the width and risk edges significant at the stream level; the pooled coverage edge is the small non-significant one in Table 1); against fallback-only it trades a small non-significant coverage gain for a width increase significant at the stream level ($+0.12\sigma$, from the A1 recursion absorbing residual mismatch) and a small extra cost. There is no free composition: each actuator class contributes the failure mode the other cannot fix, which is the empirical content of the guarantee ledger.

H.6 E3: NATURAL DEGRADATION, TWO REAL REGIME BREAKS

The scenarios above inject controlled corruptions. We complement them with two *natural*, non-injected regime breaks on ERCOT hourly system load: *Winter Storm Uri* (deployment 2021-01 through 2021-07; the February 2021 Texas freeze, event window 144 h) and the *COVID-19 demand-pattern shift* (deployment 2020-02 through 2020-08; event window 1032 h). No data is modified: the incident itself is the degradation, it hits the truth y (not an input view), the event windows are taken from public incident records, and deployment is anchored 720 h before onset so the gates monitor

through the break. These runs are deterministic (nothing is injected), so Table 6 reports one run per (stream, policy).

Three measured facts (Table 6). First, static-width policies pay at the event: static conformal and fallback-only drop to 0.729 coverage inside the Uri window (rolling minimum 0.738), while the recursion-carrying policies hold 0.88–0.89 through it; the deterministic ledger absorbing a natural break as it absorbs an injected one. (The COVID shift, by contrast, moves the level slowly enough that even static widths keep covering: event coverage 0.946: a natural regime break need not be a coverage event.) Second, the e-gates fire *zero* times on both streams: a natural regime break in the truth degrades the model and the seasonal-naive baseline together, so the fallback null (model worse than baseline) genuinely never fails persistently, and the gates' silence is correct behavior out-of-distribution, not a miss; across both natural streams there is no false structural intervention, consistent with Theorem 3.1(b) and with $M_T = 0$ in Theorem 3.3: when no *relative* degradation episode exists, the coupled bound says structural machinery must contribute nothing, and it does not. Third, as a consequence, the unified policy reduces to its deterministic ledger plus a small ungated-repair overhead (12 and 11 cost units) with no benefit over plain ACI here: on breaks that hit all pipelines symmetrically, the structural library is dead weight and only the calibration actuators earn their keep.

H.7 E4: NATURAL ASYMMETRIC INCIDENTS, REAL UPSTREAM FEED FAILURES

The two breaks above hit all pipelines symmetrically, so the gates' silence is the correct behavior: validity, but not power. We therefore add natural incidents of the opposite class: real, recorded *upstream feed failures* that hit one input view only. One deterministic rule, applied uniformly to two public benchmarks (PEMS road occupancy, 862 hourly sensors, 2016–2018; ECL electricity load, 321 hourly clients, 2016–2019), defines an incident: a sensor whose longest maximal zero-run of ≥ 100 h (starting after hour 2000) follows a clean history ($< 2\%$ zeros) and for which a healthy target sensor ($< 1\%$ zeros, pre-failure correlation ≥ 0.85) exists. The scan returns exactly two incidents: traffic sensor 181, dead from hour 16,225 to end of file (1,319 h; 1.5% zeros before), and electricity client 182, dead from hour 6,390 to end of file (19,914 h; 0.02% zeros before). One further candidate (electricity client 245) has a qualifying zero-run but no eligible target and is excluded by the rule. Nothing is injected: both outages are present in the raw public files, and the event windows are detected from the files themselves (longest zero-run). The served model is a cross-sensor nowcast ("virtual sensor"): it predicts a healthy target sensor from recent readings of the feed, the standard setup when a location's own detector is unavailable; the seasonal-naive baseline uses only the target's own history, so the feed failure degrades the model *relative* to the baseline; a genuine non-null for the fallback gate. Targets are selected deterministically on pre-deployment data only (minimizing the nowcast-vs-baseline calibration-split loss gap: sensor 502, client 253); the repair actuator imputes dead feed readings by their weekly-seasonal history; deployment is anchored 720 h before onset.

This is the power evidence the symmetric streams cannot give (Table 7). The fallback gate *fires on both incidents*. On the traffic outage, the unified policy switches to the baseline pipeline 134 h after the feed dies and serves it through the outage (fallback fraction 0.70; the gate also fired once during a pre-onset stretch where the nowcast genuinely underperformed, with the recovery gate correctly returning to the model: all firings and times are in the released raw): it reaches 0.844 event coverage at width 1.82, while the calibration-only policies must widen enormously to chase a dead model; ACI reaches only 0.751 at width 2.80, conformal PI 0.858 at width 3.00, $1.6\times$ wider than unified for a $+0.014$ event-coverage difference. On the electricity outage: longer (19,914 h) and detected faster (the gate fires 44 h after onset, a single firing, fallback fraction 0.96); the separation is starker: the unified policy holds the budget through the entire outage (0.900 event coverage at width 0.68) while conformal PI needs width 3.78 ($5.5\times$ wider) for the same event coverage (0.899), and ACI 3.80 for 0.891. Ungated fallback-only cannot recalibrate the baseline it serves: on traffic, where the static baseline width is miscalibrated for the target, it manages only 0.435 event coverage; on electricity, where the baseline happens to be well calibrated, it does well (0.932 at width 0.81) but overshoots the budget where the unified policy's recursion tracks it (0.900) at narrower width; the fallback-only success is baseline luck. The static policy collapses on both (0.051 / 0.001). Two caveats: the traffic break is severe enough that *no* policy holds the 0.90 target over the full deployment (unified 0.855, PI 0.874); both outages run to the end of their files so no post-event recovery is observable, and the scan yields no recovered (non-terminal) incident under the rule, as the only candidate with a recovered zero-run (electricity client 245) fails the target-eligibility rule, so the recovery gate's natural-data

Table 6: Natural (non-injected) regime breaks on ERCOT hourly load: Winter Storm Uri (event window 144 h, Feb 2021) and the COVID-19 demand shift (event window 1032 h, spring 2020). *cov.* = deployment coverage; *event* = coverage inside the event window; *roll-min* = minimum 168 h rolling coverage; *width* in calibration-sd units; *cost* = total intervention cost. One deterministic run per cell (nothing is injected).

	ERCOT / Storm Uri					ERCOT / COVID-19				
policy	cov.	event	roll-min	width	cost	cov.	event	roll-min	width	cost
Static SC	0.927	0.729	0.738	0.138	0	0.941	0.946	0.863	0.122	0
ACI (R1)	0.900	0.882	0.851	0.128	0	0.900	0.899	0.857	0.107	0
Conformal PI (R2)	0.900	0.889	0.863	0.133	0	0.900	0.899	0.881	0.108	0
Online CRC (R3)	0.011	0.000	0.000	0.001	0	0.011	0.000	0.000	0.001	0
e-Fallback (R6)	0.927	0.729	0.738	0.138	0	0.941	0.946	0.863	0.122	0
Debt-Repair (R4)	0.900	0.882	0.845	0.129	12	0.900	0.899	0.857	0.108	9
Unified DRC	0.900	0.882	0.845	0.129	12	0.900	0.899	0.857	0.108	9

Table 7: Natural asymmetric incidents: recorded terminal outages of PEMS sensor 181 (dead from hour 16,225; 1,319 h) and ECL client 182 (dead from hour 6,390; 19,914 h), each degrading only the cross-sensor nowcast's input view. *fires* = structural gate firings; *fb* = fraction of deployment served by fallback. One deterministic run per cell.

	traffic / sensor 181							electricity / client 182						
policy	cov.	event	roll-min	width	cost	fires	fb	cov.	event	roll-min	width	cost	fires	fb
Static SC	0.330	0.051	0.000	0.68	0	0	0.00	0.030	0.001	0.000	0.66	0	0	0.00
ACI (R1)	0.803	0.751	0.220	2.80	0	0	0.00	0.892	0.891	0.429	3.80	0	0	0.00
Conformal PI (R2)	0.874	0.858	0.702	3.00	0	0	0.00	0.898	0.899	0.661	3.78	0	0	0.00
Online CRC (R3)	0.536	0.469	0.000	2.05	0	0	0.00	0.533	0.534	0.119	3.14	0	0	0.00
e-Fallback (R6)	0.552	0.435	0.000	0.57	285	3	0.70	0.928	0.932	0.565	0.81	3974	1	0.96
Debt-Repair (R4)	0.804	0.751	0.238	2.77	47	0	0.00	0.892	0.892	0.446	3.78	279	0	0.00
Unified DRC	0.855	0.844	0.435	1.82	295	3	0.70	0.900	0.900	0.649	0.68	3985	1	0.96

evidence is limited to the traffic pre-onset round trip; and these are two deterministic runs on the two incidents a fixed public-benchmark scan yields; power evidence for the incident class the theory targets (asymmetric, persistent, detectable), not a distributional claim.

H.8 E1: ROBUSTNESS TO THE BACKBONE FAMILY

The framework's claims are about the intervention loop, not the predictor, and a second backbone family tests that. We replicate the full E1 design (9 policies $\times$ 6 streams $\times$ 3 scenarios $\times$ 10 seeds) with the linear AR backbone replaced by a shallow MLP (64×32 hidden units, same lag features, early stopping; seeded per run), everything else identical (Table 8). The qualitative structure is unchanged: under degradation the unified policy is the only one at coverage 0.900 ± 0.003 with max shortfall 0.006, holding the budget at width 0.64 while conformal PI needs 1.52 ($2.4\times$ wider; significant at the stream level), DtACI 1.48, betting-ACI 1.44, and ACI 1.43; online CRC's coverage collapses while it controls its own graded risk; and every failure mode replicates; the clean-stream repair overhead (41.9 cost units), the non-significant coverage gain over fallback-only (mirroring the linear backbone), and degraded-both, where conformal PI again beats the unified policy on coverage (0.897 vs 0.888, significant at the stream level) while the unified policy pays 306 cost units of useless structural actions. Two backbone-specific observations: the MLP's static baseline is noisier across seeds than the linear one (clean noop 0.906 ± 0.051), and for the MLP the recalibration-only policies land identically in degraded and degraded-both (they never touch the baseline view), so the second corruption is invisible to them by construction. We do not pool the two backbones: each is reported against its own backbone's runs.

H.9 E1: FOUNDATION-MODEL BACKBONES

A third and a fourth backbone family stress the claim from the opposite direction: a pretrained time-series foundation model (TSFM) is not fit to the stream at all, so its failure modes under input corruption are not those of a regression on lag features. We replicate the full E1 design ($7 \times 6 \times 3 \times 10$) with the served forecast produced by two TSFMs from distinct families (Chronos-Bolt (small) (Ansari et al., 2024) and TimesFM-2.5 (200M) (Das et al., 2024)) each as a separate full replication: zero-shot

Table 8: Second backbone family (shallow MLP), same design and budget as Table 1: pooled over six streams, 10 seeds, three scenarios. *shortfall* = mean/max of $(0.90 - \text{cov})_+$.

	clean				degraded				degraded-both			
policy	cov.	width	cost	shortfall	cov.	width	cost	shortfall	cov.	width	cost	shortfall
Static SC	0.906 $\pm$0.051	0.26	0	0.021/0.081	0.440 $\pm$0.040	0.26	0	0.460/0.527	0.440 $\pm$0.040	0.26	0	0.460/0.527
ACI (R1)	0.900 $\pm$0.001	0.25	0	0.000/0.001	0.886 $\pm$0.004	1.43	0	0.014/0.025	0.886 $\pm$0.004	1.43	0	0.014/0.025
DtACI	0.899 $\pm$0.002	0.26	0	0.001/0.005	0.896 $\pm$0.001	1.48	0	0.004/0.010	0.896 $\pm$0.001	1.48	0	0.004/0.008
Betting ACI	0.886 $\pm$0.002	0.24	0	0.014/0.017	0.879 $\pm$0.001	1.44	0	0.021/0.024	0.879 $\pm$0.001	1.44	0	0.021/0.024
Conformal PI (R2)	0.900 $\pm$0.000	0.26	0	0.000/0.000	0.897 $\pm$0.000	1.52	0	0.003/0.004	0.897 $\pm$0.000	1.52	0	0.003/0.004
Online CRC (R3)	0.015 $\pm$0.006	0.00	0	0.885/0.892	0.414 $\pm$0.041	0.79	0	0.486/0.618	0.414 $\pm$0.041	0.79	0	0.486/0.618
e-Fallback (R6)	0.906 $\pm$0.051	0.26	0	0.021/0.081	0.874 $\pm$0.062	0.52	863	0.038/0.139	0.486 $\pm$0.055	0.34	262	0.414/0.513
Debt-Repair (R4)	0.900 $\pm$0.001	0.25	42	0.000/0.002	0.886 $\pm$0.004	1.43	59	0.014/0.025	0.886 $\pm$0.004	1.43	59	0.014/0.025
Unified DRC	0.900 $\pm$0.001	0.25	42	0.000/0.002	0.900 $\pm$0.003	0.64	884	0.001/0.006	0.888 $\pm$0.002	1.37	306	0.012/0.017

median forecasts from 512-step contexts of the (possibly degraded) input view, in non-overlapping 24-step blocks, weights frozen; the models never see the streams at training time (Table 9; Table 12). Chronos-Bolt carries the running prose; the TimesFM replication is reported at the end of this subsection.

On the injected-noise scenario the width frontier does not appear, because the backbone itself absorbs the corruption. A zero-shot median forecast from a 512-step context is intrinsically robust to i.i.d. input noise: static conformal drops only to 0.786 (linear backbone: far lower), every recursion-carrying policy sits at 0.884–0.899 at width 0.52–0.54 (just $1.35\times$ its clean width; DtACI and betting-ACI land inside the same cluster, 0.898/0.53 and 0.884/0.52, marginally narrower than the unified policy at slightly lower coverage) and the fallback gate correctly declines to fire in 59 of the 60 degraded stream–seed configurations (it fires once, on ETTh1 seed 0, spending 16% of that run in fallback), since the degraded TSFM still beats the seasonal-naive baseline, so the fallback null genuinely holds. The unified policy reduces to ACI plus $\sim$80 units of repair overhead. This is the framework behaving as specified under a backbone that removes the relative-degradation signal (the same mechanism as the symmetric natural breaks (Appendix H.6)) and it delimits the empirical claim: the frontier belongs to deployments where degradation actually defeats the served model, not to every deployment. Two further backbone-specific observations: online CRC collapses catastrophically here (clean coverage 0.072), because its graded-risk recursion tracks an error scale that the zero-shot quantile geometry does not supply; and the same collapse appears on the natural ERCOT replays (Uri coverage 0.373), where the gates again correctly do not fire.

The frontier reappears under the TSFM at the point the theory locates it: when the incident is structural rather than statistical. On the real feed failures (the TSFM forecasts a target-scale affine mapping of the feed: a virtual-sensor nowcast; Chronos is univariate, so the cross-sensor structure enters through the mapping, not the model) a dead feed defeats any forecaster reading it, foundation model or not: the gate fires on both incidents (3 firings / 1 firing; fallback fractions 0.74 / 0.96), and the electricity outage reproduces the full separation; unified 0.901 event coverage at width 0.68 vs conformal PI 0.898 at 4.05 ($5.9\times$ wider) and ACI 0.891 at 4.05; on traffic, unified reaches 0.846 event coverage at width 1.83 vs PI 0.871 at 2.92 ($1.6\times$). The second family replicates the pattern on the same incidents (Table 13): under TimesFM-2.5 the gate fires on both (fallback fractions 0.73 / 0.98), the electricity outage again reproduces the full separation (unified 0.900 event coverage at width 0.68 vs conformal PI 0.898 at 4.05 ($6.0\times$ wider)) and on traffic, unified 0.844 at 1.83 vs PI 0.869 at 2.90 ($1.6\times$). That the two families' E4 numbers nearly coincide (while their E1/E5 numbers differ systematically) is itself structural: through the outage the gated policies are served 73–98% of the time by the *same* seasonal-naive fallback, and the recalibration-only widths are dominated by the dead feed's mapping residuals; the incident drowns the backbone difference, which is precisely the point of the incident. Two models from two TSFM families on these streams: evidence about where the intervention loop earns its keep within these backbones.

The TimesFM-2.5 replication (E1). The second family reproduces the injected-noise null finding independently (Table 12): static conformal drops only to 0.792 under degradation, every recursion-carrying policy sits at 0.884–0.899 at width 0.53–0.55 ($1.38\times$ its clean width of 0.40 (Chronos-Bolt: $1.35\times$); DtACI 0.898/0.54, betting-ACI 0.884/0.53) the fallback gate produces zero firings across *all* 120 degraded stream–seed configurations (under Chronos-Bolt it fired once), and the unified policy again reduces to ACI plus $\sim$76 units of repair overhead. The online-CRC collapse is, if anything, sharper (clean coverage 0.076). The two families were run as separate replications and never pooled; that both, independently, absorb the injected corruption while both fail on the real dead feeds is precisely the boundary the theory draws.

Table 9: Third backbone family (Chronos-Bolt TSFM, zero-shot), same design and budget as Table 1: pooled over six streams, 10 seeds, three scenarios. *shortfall* = mean/max of $(0.90 - \text{cov})_+$.

	clean				degraded				degraded-both			
policy	cov.	width	cost	shortfall	cov.	width	cost	shortfall	cov.	width	cost	shortfall
Static SC	0.903 $\pm$0.037	0.41	0	0.015/0.046	0.786 $\pm$0.073	0.41	0	0.114/0.235	0.786 $\pm$0.073	0.41	0	0.114/0.235
ACI (R1)	0.900 $\pm$0.001	0.39	0	0.000/0.001	0.898 $\pm$0.001	0.54	0	0.002/0.005	0.898 $\pm$0.001	0.54	0	0.002/0.005
DtACI	0.898 $\pm$0.000	0.39	0	0.002/0.008	0.898 $\pm$0.001	0.53	0	0.002/0.009	0.897 $\pm$0.001	0.53	0	0.003/0.006
Betting ACI	0.886 $\pm$0.002	0.37	0	0.014/0.016	0.884 $\pm$0.001	0.52	0	0.016/0.019	0.884 $\pm$0.001	0.52	0	0.016/0.019
Conformal PI (R2)	0.900 $\pm$0.000	0.39	0	0.000/0.000	0.899 $\pm$0.000	0.54	0	0.001/0.002	0.899 $\pm$0.000	0.54	0	0.001/0.002
Online CRC (R3)	0.072 $\pm$0.115	0.04	0	0.828/0.885	0.144 $\pm$0.116	0.09	0	0.756/0.877	0.144 $\pm$0.116	0.09	0	0.756/0.877
e-Fallback (R6)	0.903 $\pm$0.037	0.41	0	0.015/0.046	0.787 $\pm$0.073	0.41	3	0.113/0.235	0.786 $\pm$0.073	0.41	0	0.114/0.235
Debt-Repair (R4)	0.900 $\pm$0.001	0.39	72	0.000/0.001	0.898 $\pm$0.001	0.54	77	0.002/0.005	0.898 $\pm$0.001	0.54	77	0.002/0.005
Unified DRC	0.900 $\pm$0.001	0.39	72	0.000/0.001	0.898 $\pm$0.001	0.54	80	0.002/0.005	0.898 $\pm$0.001	0.54	77	0.002/0.005

Table 10: TSFM backbone, natural symmetric breaks (same design as Table 6). One deterministic run per cell.

	ERCOT / Storm Uri					ERCOT / COVID-19				
policy	cov.	event	roll-min	width	cost	cov.	event	roll-min	width	cost
Static SC	0.922	0.681	0.565	0.784	0	0.914	0.881	0.750	0.620	0
ACI (R1)	0.904	0.736	0.571	0.752	0	0.899	0.881	0.774	0.575	0
Conformal PI (R2)	0.902	0.799	0.661	0.756	0	0.899	0.890	0.804	0.579	0
Online CRC (R3)	0.373	0.646	0.000	0.172	0	0.217	0.114	0.000	0.083	0
e-Fallback (R6)	0.922	0.681	0.565	0.784	0	0.914	0.881	0.750	0.620	0
Debt-Repair (R4)	0.904	0.743	0.577	0.747	59	0.899	0.881	0.768	0.574	52
Unified DRC	0.904	0.743	0.577	0.747	59	0.899	0.881	0.768	0.574	52

Table 11: TSFM backbone, natural asymmetric feed failures (same incidents as Table 7; virtual-sensor nowcast). One deterministic run per cell.

	traffic / sensor 181							electricity / client 182						
policy	cov.	event	roll-min	width	cost	fires	fb	cov.	event	roll-min	width	cost	fires	fb
Static SC	0.330	0.051	0.000	0.74	0	0	0.00	0.029	0.000	0.000	0.66	0	0	0.00
ACI (R1)	0.810	0.762	0.375	2.78	0	0	0.00	0.891	0.891	0.196	4.05	0	0	0.00
Conformal PI (R2)	0.881	0.871	0.714	2.92	0	0	0.00	0.898	0.898	0.619	4.05	0	0	0.00
Online CRC (R3)	0.552	0.489	0.030	2.01	0	0	0.00	0.534	0.535	0.095	3.40	0	0	0.00
e-Fallback (R6)	0.560	0.456	0.000	0.58	303	3	0.74	0.928	0.932	0.565	0.81	3979	1	0.96
Debt-Repair (R4)	0.811	0.763	0.369	2.75	46	0	0.00	0.891	0.891	0.232	4.03	276	0	0.00
Unified DRC	0.855	0.846	0.411	1.83	311	3	0.74	0.900	0.901	0.649	0.68	3991	1	0.96

Table 12: Fourth backbone family (TimesFM-2.5 TSFM, zero-shot), same design and budget as Table 1: pooled over six streams, 10 seeds, three scenarios.

	clean				degraded				degraded-both			
policy	cov.	width	cost	shortfall	cov.	width	cost	shortfall	cov.	width	cost	shortfall
Static SC	0.910 $\pm$0.037	0.42	0	0.012/0.037	0.792 $\pm$0.071	0.42	0	0.108/0.226	0.792 $\pm$0.071	0.42	0	0.108/0.226
ACI (R1)	0.901 $\pm$0.001	0.40	0	0.000/0.001	0.899 $\pm$0.001	0.55	0	0.002/0.006	0.899 $\pm$0.001	0.55	0	0.002/0.006
DtACI	0.898 $\pm$0.001	0.39	0	0.003/0.007	0.898 $\pm$0.001	0.54	0	0.002/0.007	0.898 $\pm$0.001	0.54	0	0.002/0.006
Betting ACI	0.886 $\pm$0.002	0.38	0	0.014/0.016	0.884 $\pm$0.002	0.53	0	0.016/0.020	0.884 $\pm$0.002	0.53	0	0.016/0.020
Conformal PI (R2)	0.900 $\pm$0.000	0.40	0	0.000/0.000	0.899 $\pm$0.001	0.55	0	0.001/0.002	0.899 $\pm$0.001	0.55	0	0.001/0.002
Online CRC (R3)	0.076 $\pm$0.124	0.04	0	0.824/0.883	0.157 $\pm$0.122	0.10	0	0.743/0.879	0.157 $\pm$0.122	0.10	0	0.743/0.879
e-Fallback (R6)	0.910 $\pm$0.037	0.42	0	0.012/0.037	0.792 $\pm$0.071	0.42	0	0.108/0.226	0.792 $\pm$0.071	0.42	0	0.108/0.226
Debt-Repair (R4)	0.901 $\pm$0.001	0.40	71	0.000/0.001	0.899 $\pm$0.001	0.55	77	0.002/0.006	0.899 $\pm$0.001	0.55	77	0.002/0.006
Unified DRC	0.901 $\pm$0.001	0.40	71	0.000/0.001	0.899 $\pm$0.001	0.55	77	0.002/0.006	0.899 $\pm$0.001	0.55	77	0.002/0.006

Table 13: Second TSFM family (TimesFM-2.5), natural asymmetric feed failures (same incidents and virtual-sensor protocol as Table 11). One deterministic run per cell.

	traffic / sensor 181							electricity / client 182						
policy	cov.	event	roll-min	width	cost	fires	fb	cov.	event	roll-min	width	cost	fires	fb
Static SC	0.334	0.050	0.000	0.77	0	0	0.00	0.029	0.001	0.000	0.67	0	0	0.00
ACI (R1)	0.811	0.760	0.369	2.77	0	0	0.00	0.891	0.891	0.220	4.05	0	0	0.00
Conformal PI (R2)	0.881	0.869	0.708	2.90	0	0	0.00	0.898	0.898	0.625	4.05	0	0	0.00
Online CRC (R3)	0.549	0.489	0.030	2.01	0	0	0.00	0.532	0.535	0.077	3.40	0	0	0.00
e-Fallback (R6)	0.565	0.453	0.000	0.59	299	3	0.73	0.930	0.933	0.565	0.81	4038	1	0.98
Debt-Repair (R4)	0.811	0.760	0.375	2.74	45	0	0.00	0.891	0.891	0.244	4.03	271	0	0.00
Unified DRC	0.855	0.844	0.423	1.83	305	3	0.73	0.900	0.900	0.649	0.68	4044	1	0.98

H.10 E5: DECISION LAYER, TWO OPERATIONAL TASKS

The decision experiments route every intervention-library policy's *served intervals* (unchanged, from the same E1 runs' protocol; the two added calibration baselines are E1-only and do not enter the structural/downstream layers) through the two decision interfaces of Section 4 (Table 14). *Newsvendor*: hourly order z_t at the ρ-quantile point of the served interval ($\rho = b/(h+b) \approx 0.94$, overage $h = 1$, underage $b = 15.6$; constants monetized from a storage-holding context), realized loss $h(z_t - y_t)_+ + b(y_t - z_t)_+$ (reported untruncated; realized losses stay finite on these streams, but the theory prices the tail through B_{dec}; Proposition F.7 is about this term), certificate $R_t = \frac{hb}{h+b}|C_t|$, the width-linear in-set term of Corollary F.3 (the full minimax certificate adds the width-independent constant αB_{dec}, identical across policies, so we track the term that separates them). *Storage arbitrage*: a unit battery (round-trip $\eta = 0.9$) charges/discharges once daily at the hours maximizing the certified profit floor $\eta L_s - U_b$ computed from the served bands; reported are the certified floor, realized profit, and the fraction of days realized profit falls below the certified floor. One caveat: the served bands are marginally calibrated per step, while the daily floor event intersects two coordinates, so the E5 floors carry no day-level guarantee and the violation rates below are the unadjusted measurements; the block ledger of Theorem F.10, which does control this rate, is tested separately below (E5c).

The width frontier converts into a decision frontier (Figure 3), as Corollary F.3 says it must. Under degradation (pooled over the applicable streams, 10 seeds): the unified policy attains the lowest realized newsvendor loss of all seven intervention-library policies (0.93 vs conformal PI 2.16, ACI 2.20, static 5.74) *and* the certificate-violation rate closest to the budget (0.100, i.e. α to the third decimal); the recalibration-only policies pay their extra width twice, once as certificate (R_t: PI 3.30 vs unified 1.21) and once as realized loss, while fallback-only gets loss 1.06 but violates its (narrow, static) certificates at 0.128. On storage, the same ordering appears in the trustworthiness of the floors: the unified policy's certified floors are violated on 5.9% of days (fallback-only 10.9%, PI 37.4%, ACI 38.5%, static 59.9%) at the top realized profit tier (0.559 per day; marginally below fallback-only's 0.564, which buys its profit with $1.8\times$ the floor violations); wide, well-calibrated bands make weak promises, narrow stale bands make false ones, and only the gated structural policies make narrow promises that hold. Online CRC is the instructive extreme: controlling its graded functional, it violates newsvendor certificates at 0.579 and storage floors at 0.810; Proposition F.8 in the field: a policy can look controlled on one ledger and be decision-unreliable on another. On the negative side: on clean streams the decision layer separates nothing (all calibrated policies sit at newsvendor loss 0.42–0.46, so the unified policy's repair overhead buys nothing); under degradation the unified policy's newsvendor-violation edge over fallback-only is not significant at the stream level (echoing the coverage result); and in degraded-both the picture is mixed as in E1; unified wins newsvendor loss over PI (significant at the stream level) but PI attains the lower certificate-violation rate (0.103 vs 0.113, significant at the stream level). The backbone replications appear in Tables 16 to 18.

Decision layer under the TSFM backbones. Routing the TSFM-served intervals of Appendix H.9 through the same two interfaces (same constants, full deployment segment) extends the E1 TSFM finding to the decision layer, in both directions. Under injected degradation the decision frontier vanishes together with the width frontier, as Corollary F.3 predicts: all recursion-carrying policies are decision-indistinguishable (newsvendor loss 0.75–0.77 across ACI, PI, debt-repair, and unified; certificate violations 0.101–0.102; storage floor violations 0.017–0.019), the unified policy's structural machinery buys nothing, and the only separations are static conformal and e-fallback paying the ledger transient (loss 0.98, violations 0.213–0.214; the gate correctly never fires, so e-fallback coincides with its static base) and online CRC being decision-unreliable everywhere (newsvendor certificate violations 0.856, floor violations 0.479). No width gap, no loss gap: the corollary read in reverse, and the TSFM decision headline is a negative result. The TimesFM-2.5 replication reproduces the vanishing independently: recursion-carrying policies at newsvendor loss 0.76–0.79 with certificate violations at 0.101 and storage floor violations 0.023–0.026; static and e-fallback (no firing, so they coincide) at loss 0.99 and violations 0.208; online CRC decision-unreliable again (newsvendor violations 0.843, floor violations 0.482). Per-task tables for both TSFM families: Tables 17 and 18.

E5c: the day-level joint ledger in the field. Running the storage task with the block recursion of Theorem F.10 ($\gamma_D = 0.05$, $\alpha_D = 0.1$, 291 trading days per run, five price streams, linear and MLP backbones; Table 15) turns the E5 caveat into a controlled quantity. Uncontrolled day-level floor violation rates under degradation range from 0.059 (unified) to 0.810 (online CRC): a $14\times$ spread across policies sharing one step-level budget, the measured content of the caveat. With the block

Table 14: Decision layer under degradation (linear backbone; pooled, 10 seeds): newsvendor (six streams) and storage arbitrage (five price streams). *loss* = realized decision loss; *cert* = mean served certificate R_t; *viol* = certificate-violation rate / fraction of days below the certified floor; *floor* / *profit* in σ-units per day.

	newsvendor			storage arbitrage		
policy	loss	cert	viol	floor	profit	viol
Static SC	5.74	0.53	0.541	3.02	0.300	0.599
ACI (R1)	2.20	3.09	0.116	0.48	0.298	0.385
Conformal PI (R2)	2.16	3.30	0.103	0.36	0.284	0.374
Online CRC (R3)	3.13	1.86	0.579	1.71	0.299	0.810
e-Fallback (R6)	1.06	0.99	0.128	-0.12	0.564	0.109
Debt-Repair (R4)	2.19	3.08	0.116	0.47	0.295	0.385
Unified DRC	0.93	1.21	0.100	-0.40	0.559	0.059

Table 15: E5c: day-level joint floor violations with the block ledger of Theorem F.10 ($\alpha_D = 0.1$, $\gamma_D = 0.05$, $D = 291$ days; pooled over five price streams, 10 seeds). *viol* = measured day-level violation rate with the block recursion; *unctrl* = the same rate without it (the E5 caveat, measured); *floor* = mean certified floor (σ-units).

	linear						MLP					
	clean			degraded			clean			degraded		
policy	viol	unctrl	floor	viol	unctrl	floor	viol	unctrl	floor	viol	unctrl	floor
Static SC	0.104	0.084	+0.32	0.306	0.599	+0.35	0.105	0.102	+0.37	0.283	0.669	+0.22
ACI (R1)	0.100	0.076	+0.30	0.164	0.385	-0.44	0.102	0.095	+0.36	0.159	0.408	-0.42
Conformal PI (R2)	0.100	0.079	+0.29	0.161	0.374	-0.42	0.102	0.091	+0.35	0.158	0.395	-0.40
Online CRC (R3)	0.124	0.650	+0.36	0.216	0.810	-0.16	0.124	0.847	+0.39	0.212	0.909	-0.20
e-Fallback (R6)	0.104	0.084	+0.32	0.100	0.109	-0.07	0.105	0.102	+0.37	0.100	0.095	-0.12
Debt-Repair (R4)	0.099	0.072	+0.30	0.163	0.385	-0.44	0.102	0.090	+0.36	0.158	0.408	-0.42
Unified DRC	0.099	0.072	+0.30	0.089	0.059	-0.30	0.102	0.090	+0.36	0.089	0.076	-0.33

ledger every policy is pulled toward α_D: on clean streams all seven intervention-library policies land at 0.099–0.124; under degradation the gated policies sit at 0.089–0.100 while the ungated ones still show 0.161–0.306 at $D = 291$, and this residual is not a violation of the theorem but its content: the deviation equals $\lambda_{D+1}/(\gamma_D D)$ *exactly* (the telescoped ledger identity; e.g. static conformal's pooled excess 0.206 matches $\lambda_{D+1}/(\gamma_D \cdot 291) = 0.206$ to six decimals in the raw), i.e. the finite-D transient of a λ still ramping against an ongoing degradation, vanishing at the theorem's $O(1/\gamma_D D)$ rate. The price is observable and two-sided, as the recursion is: policies whose floors were false pay tightness (ACI's mean floor drops from $+0.48$ to -0.44 under degradation), and the one policy that was *over*-delivering (unified, uncontrolled rate $0.059 < \alpha_D$) has its unused budget spent back on tighter floors (λ drifts negative, rate rises to 0.089). Realized profits are unchanged to the digit: for this certificate the inflation is a schedule-independent offset (Appendix F.4), so the ledger reallocates trust while revenue is untouched, by construction. What survives of the policy ranking: the gated policies reach the budget with the smallest transient because they started nearest it; the day-level ledger equalizes long-run rates by design, so the remaining separation axis is the one Theorem F.10 prices; floor tightness at a controlled violation rate. A (γ_D, α_D) sensitivity scan on the same logged bands ($\gamma_D \in \{0.02, 0.05, 0.1, 0.2\}$, $\alpha_D \in \{0.05, 0.1\}$, degraded scenario) confirms the transient scales as the theorem says it must: the worst pooled excess over α_D shrinks monotonically in γ_D; static conformal's rate at $\alpha_D = 0.1$ falls from 0.505 ($\gamma_D = 0.02$) through 0.306 (0.05) and 0.207 (0.1) to 0.154 (0.2), i.e. inversely in γ_D per the $O(1/\gamma_D D)$ bound; while the gated policies stay within 0.025 of α_D at every grid point and both α_D levels are tracked with the same geometry; larger γ_D buys a shorter transient at the price of coarser λ steps, exactly the trade Equation (2) exposes.

H.11 E6: THE META-POLICY IN THE FIELD

E6 runs the sampled interleaved-Hedge meta-policy of Theorem G.2 over the full seven-policy library on the per-step truncated newsvendor losses of the E1/E5 protocol ($K = 7$, $H = 1$, $B = 20$ in σ-units, $\delta = 0.05$; the truncation level is set at roughly $b \times 1.3\sigma$; above the untruncated loss on 99.91% of clean-scenario steps, though 2.6–3.5% of injected-degradation steps do clip (measured over the full linear design); since the truncation defines the loss identically for the meta-policy and all experts, the regret comparison is internally consistent, and B trades bound tightness against

Table 16: Decision layer under degradation, MLP backbone (same interfaces and constants as Table 14).

	newsvendor			storage arbitrage		
policy	loss	cert	viol	floor	profit	viol
Static SC	4.41	0.48	0.560	2.73	0.306	0.669
ACI (R1)	1.77	2.69	0.114	0.50	0.304	0.408
Conformal PI (R2)	1.77	2.85	0.103	0.41	0.291	0.395
Online CRC (R3)	2.58	1.48	0.586	1.73	0.305	0.909
e-Fallback (R6)	1.03	0.97	0.126	-0.05	0.598	0.095
Debt-Repair (R4)	1.77	2.69	0.114	0.50	0.302	0.408
Unified DRC	0.90	1.21	0.100	-0.34	0.594	0.076

Table 17: Decision layer under degradation, Chronos-Bolt TSFM backbone (same interfaces and constants as Table 14).

	newsvendor			storage arbitrage		
policy	loss	cert	viol	floor	profit	viol
Static SC	0.98	0.76	0.214	-0.18	0.519	0.057
ACI (R1)	0.77	1.02	0.102	-0.45	0.514	0.017
Conformal PI (R2)	0.75	1.01	0.101	-0.44	0.501	0.018
Online CRC (R3)	1.74	0.17	0.856	0.41	0.520	0.479
e-Fallback (R6)	0.98	0.76	0.213	-0.18	0.519	0.057
Debt-Repair (R4)	0.78	1.02	0.102	-0.43	0.505	0.020
Unified DRC	0.78	1.02	0.102	-0.43	0.505	0.020

Table 18: Decision layer under degradation, TimesFM-2.5 TSFM backbone (same interfaces and constants as Table 14).

	newsvendor			storage arbitrage		
policy	loss	cert	viol	floor	profit	viol
Static SC	0.99	0.79	0.208	-0.19	0.504	0.079
ACI (R1)	0.79	1.04	0.101	-0.46	0.492	0.023
Conformal PI (R2)	0.76	1.03	0.101	-0.45	0.475	0.025
Online CRC (R3)	1.77	0.18	0.843	0.41	0.501	0.482
e-Fallback (R6)	0.99	0.79	0.208	-0.19	0.504	0.079
Debt-Repair (R4)	0.79	1.04	0.101	-0.45	0.482	0.027
Unified DRC	0.79	1.04	0.101	-0.45	0.482	0.027

tail sensitivity rather than being a quantile claim; six streams $\times$ three scenarios $\times$ 10 seeds per backbone, $T = 7000$ steps; Table 19). By Assumption G.1 the seven in-silico expert trajectories *are* the seven E1 runs, so E6 adds only the meta layer and its sampling randomness. Three findings. First, the premise of the theorem is real: the best expert in hindsight changes with the scenario (on clean streams it is an ungated recursion-carrying policy (conformal PI or ACI; in all $60/60$ runs per backbone), under asymmetric degradation it is the unified policy ($41/60$ linear, $50/60$ MLP), and in degraded-both the honors split (unified/PI/ACI/debt-repair), so no fixed library member dominates, which is the gap the meta-policy closes. Second, the meta-policy tracks it: final average regret to the best expert is 0.02–0.10 per step across scenarios and backbones, uniformly below the anytime bound of Theorem G.2 in all 360 runs; though the bound is loose in this regime (its smallest value at $T = 7000$ is 2.12, an order of magnitude above the worst measured regret of 0.149; the slack is the price of the $B = 20$ truncation level and the uniform-in-T spending, not of the data). Third, the negative results the milestone predicted: on clean streams the experts are nearly indistinguishable, so the meta-policy never concentrates (largest pick fraction ≈ 0.20, barely above uniform $1/7$) and pays an exploration transient of $+0.02$ per step over the best expert; a real cost with no compensating certificate, since (as Appendix G.1 states) the meta level holds a regret guarantee, not an e-gated intervention ledger.

Table 19: E6: sampled interleaved-Hedge meta-policy over the seven-policy library (per-step truncated newsvendor loss, σ-units; pooled over six streams, 10 seeds, $T = 7000$). *meta* = meta-policy average loss; *best expert* = best fixed library member in hindsight (modal winner, count); *regret / bound* = mean final regret / smallest theorem bound at T.

	linear			MLP		
scenario	meta	best expert	regret / bound	meta	best expert	regret / bound
clean	0.435	0.414 (conformal PI, 30/60)	0.021 / 2.12	0.400	0.379 (ACI, 40/60)	0.021 / 2.12
degraded	0.952	0.867 (unified, 41/60)	0.084 / 2.12	0.953	0.853 (unified, 50/60)	0.100 / 2.12
degraded-both	1.967	1.891 (unified, 32/60)	0.075 / 2.12	1.735	1.683 (conformal PI, 22/60)	0.052 / 2.12

Table 20: E7: tracking regret to the $S{=}1$ oracle (per-step, σ-units; mean over six streams $\times$ 10 seeds, $T = 7000$) for GCL, the ungated windowed leader, and Hedge; adopted switches per run and median post-onset confirmation delay.

backbone	scenario	GCL	leader	Hedge	adopt/run	delay
linear	clean	0.034±0.020	0.002±0.002	0.021±0.008	0.0	–
linear	degraded	0.215±0.041	-0.002±0.011	0.087±0.025	1.2	253
linear	degraded-both	0.240±0.146	-0.005±0.015	0.078±0.043	0.8	232
mlp	clean	0.032±0.017	0.002±0.002	0.021±0.006	0.0	–
mlp	degraded	0.255±0.085	0.002±0.016	0.106±0.017	1.7	234
mlp	degraded-both	0.183±0.139	-0.006±0.015	0.054±0.033	1.0	229

H.12 E7: GATE-CONFIRMED LEADER IN THE FIELD

E7 runs the GCL meta-policy of Theorem G.3 on the same in-silico loss matrices as E6 ($K = 7$, $H = 1$, $B = 20$, nomination and gate window $W = 500$, $\alpha_{\text{meta}} = 0.05$; Table 20), against three references on identical losses: the ungated windowed leader (GCL's own nomination served directly, no certificate), the Hedge meta-policy of E6, and the $S{=}1$ oracle comparator (best expert before the onset, best expert after; the natural switching comparator, which the fixed-expert comparator of Theorem G.2 cannot represent). Three findings, one per layer of the theorem. First, the certificate is real and free of false alarms in the field: on clean streams the gate adopted *zero* switches in all 120 runs across both backbones (the measured content of part (a)) while the ungated leader switches freely (and cheaply, since clean experts are near-tied; its freedom is unearned rather than punished here). Second, adoption works where it should, but not always: under injected degradation the gate confirmed at least one switch in 50–59 of 60 runs per scenario and backbone, with median post-onset confirmation delay 229–253 steps, inside the $W + H + 1$ allowance of part (b); in the remaining runs the gate never accumulated the evidence within its spending schedule; the theorem's detectability caveat measured in the field, since nothing guarantees the injected gap clears the margin on every stream and seed. The margin is worth stating: at the E7 parameters the nomination-margin condition of part (b) demands $\Delta \geq B\sqrt{2\ln(2KT(T{+}1)/\delta)/W} \approx 0.31B \approx 6.1$ σ-units per step ($W = 500$, $T = 7000$, $K = 7$, $\delta = 0.05$); the theoretically trackable class at these parameters is nearly empty, and the measured confirmations are driven by the mixture-betting growth of (†), not by the conservative Azuma-union margin; part (b)'s sufficient condition is loose precisely in the regime in which the gate is empirically strong. (Instantiating the remaining confidence terms the same way: with $S = 1$ and a single power slice $\beta_{\text{det}} = 0.05$, the joint confidence of part (b) is $1 - \alpha_{\text{meta}} - \beta_{\text{det}} - \delta = 0.85$.) The window is not a tuned constant: rerunning E7 (linear) with $W \in \{250, 500, 1000\}$ on the same loss matrices leaves the qualitative picture unchanged; zero clean adoptions in all 60 runs at every window; 50–56 of 60 degraded runs confirm, with median delays 209–253 steps, each inside its own $W + H + 1$ allowance (at $W = 250$ by 248.5 vs. 252); and degraded tracking regret *improves* mildly with larger W ($0.31 \to 0.20$), because in this regime the extra detections outweigh the longer permitted delay. Third, the price of certification: GCL's tracking regret to the $S{=}1$ oracle (0.18–0.26 under degradation, both backbones) exceeds both the ungated leader's (≈ 0) and Hedge's (0.05–0.11), because Hedge hedges through the transition while GCL holds a certified incumbent until the gate fires, and pays again on the undetected runs. The measured ordering (ungated $<$ Hedge $<$ GCL) is the empirical face of Proposition G.4: the excess is what an α_{meta}-certificate on every switch costs, and the deployment chooses whether each meta switch must be defensible or merely profitable.

Table 21: E9: MS-GCL (dyadic grid $\{125, 250, 500, 1000\}$, budget $\alpha_{\text{meta}}/4$ per scale) vs. single-scale GCL ($W = 500$) on identical loss matrices. Runs with ≥ 1 adopted switch, median post-onset confirmation delay, and tracking regret to the S=1 oracle.

backbone	scenario	Runs adopting		Median delay		Track. regret	
		MS	GCL	MS	GCL	MS	GCL
linear	`clean`	0/60	0/60	—	—	0.034	0.034
linear	`degraded`	52/60	50/60	270.5	253.0	0.207	0.215
linear	`degraded_big`	60/60	60/60	124.0	144.5	0.146	0.157
MLP	`clean`	0/60	0/60	—	—	0.032	0.032
MLP	`degraded`	60/60	59/60	243.0	234.0	0.199	0.255
MLP	`degraded_big`	60/60	60/60	162.5	163.0	0.168	0.174

E9: the multi-scale grid, measured. E9 runs MS-GCL of Theorem G.6 against single-scale GCL on identical loss matrices (dyadic grid $W_i \in \{125, 250, 500, 1000\}$, per-scale budget $\alpha_{\text{meta}}/4$; single scale $W = 500$ as E7; Table 21), on clean, degraded, and a *designed* large-gap scenario degraded-big (E1's noise operator at 6.0 standard deviations instead of 1.5, model view only); a designed scenario (the E8b convention), used to place the post-onset gap far above the W=500 detection level so that the gap-adaptive delay claim of part (b) has something to bite on; it is protocol coverage, not a natural-frequency claim. Findings, per layer. *Validity:* on clean streams MS-GCL adopted zero switches in all 120 runs across backbones, despite running four gates instead of one; part (a)'s content at the split budget. *Gap-adaptive delay:* median post-onset confirmation delay shrinks with the gap under both protocols (linear: 270.5 steps at 1.5σ vs. 124.0 at 6σ for MS-GCL; 253.0 vs. 144.5 for single-scale), and MS-GCL confirms earlier than single-scale on the large-gap scenario while adopting in all $60/60$ runs on both backbones (124.0 vs. 144.5 linear; 162.5 vs. 163.0 MLP; a real but marginal edge there). *A null result:* the empirical gains over single-scale GCL are modest ($144.5 \to 124.0$ median delay; tracking regret $0.157 \to 0.146$, linear degraded-big), because the single-scale e-process already fires early inside its window when the gap is large (the mixture-betting wealth grows at the actual gap, not at Δ) so much of the worst-case B/Δ factor that Theorem G.6 removes is a property of the single-scale *bound*, not always of its field behavior. The theorem's content is the worst case (the bound now matches Proposition G.4 on the homogeneous class at every gap level); the field content at these parameters is a measured but modest improvement, and most MS-GCL confirmations still land at the larger scales (per-scale firing counts in the raw).

E11: candidate locking, measured. E11 runs SR-GCL of Theorem G.7 against MS-GCL and single-scale GCL on identical loss matrices (same grid and budgets as E9; Table 22), each from two starts: the default incumbent and a *designed* worst start; the ex-post worst post-onset policy (the E8b convention): protocol coverage of the "arbitrarily bad incumbent" scenario the theorem prices, not a deployable initialization (on degraded-big the ex-post worst *is* the default start, so the two coincide there). Findings, per layer. *Validity:* on clean streams SR-GCL adopted zero switches in all 240 runs (both backbones, both starts); part (a)'s content under locking. *Where locking helps:* on degraded (1.5σ, a near-tied post-onset cluster) SR-GCL confirms in $60/60$ runs on both backbones where MS-GCL manages 52–60, with *shorter* median post-onset delay (165.0 vs. 270.5 linear; 198.5 vs. 243.0 MLP): oscillating nominations can no longer reset a running confirmation, which is the first blocker the protocol change removes. *A cost:* on degraded-big (one dominant expert) locking *hurts*; median delay 218.0 vs. 124.0 (linear; 215.0 vs. 162.5 MLP), more total adoptions (81–97 vs. 68), and a served loss behind MS-GCL (1.126 vs. 0.995 linear); because a locked coarse scale must wait for its own window boundary to refresh a stale candidate while unlocked nomination re-points within steps; this is the locking latency that Theorem G.7's $O(BW_0)$-per-phase terms price, visible in the field; Figure 4 shows both sides of this trade at the run level. The theorem's content is the worst case over level-stable heterogeneous segments (where unlocked MS-GCL has no bound at all); the field content is a measured trade: locking buys confirmation on near-tied clusters at the price of latency on dominant-gap streams.

H.13 E8: THE COMPOSITE LEDGER, MEASURED

E8 measures Theorem G.8 on the served trajectory (Table 23). Both budget conventions are run in full: all seven policies re-executed at per-policy budget α_{S} (deployment, composite level $K\alpha_{\text{S}} + \alpha_{\text{meta}} =$

Table 22: E11: SR-GCL (candidate locking; Theorem G.7) vs. MS-GCL and single-scale GCL on identical loss matrices, from the default and the designed worst start. Runs with ≥ 1 adopted switch, median post-onset confirmation delay (convention: each run contributes its *first* adopted switch, kept only if post-onset; runs whose first adoption is pre-onset are excluded from the median), and mean served loss.

backbone	scenario	start	Runs adopting		Median delay		Served loss		
			SR	MS	SR	MS	SR	MS	GCL
linear	clean	default	0/60	0/60	—	—	0.448	0.448	0.448
linear	clean	worst	0/60	0/60	—	—	0.448	0.448	0.448
linear	degraded	default	60/60	52/60	165.0	270.5	1.099	1.072	1.080
linear	degraded	worst	60/60	60/60	160.0	263.5	1.074	1.039	1.047
linear	degraded_big	default	60/60	60/60	218.0	124.0	1.126	0.995	1.006
linear	degraded_big	worst	60/60	60/60	218.0	124.0	1.126	0.995	1.006
MLP	clean	default	0/60	0/60	—	—	0.411	0.411	0.411
MLP	clean	worst	0/60	0/60	—	—	0.411	0.411	0.411
MLP	degraded	default	60/60	60/60	198.5	243.0	1.062	1.046	1.103
MLP	degraded	worst	60/60	60/60	195.0	239.0	1.065	1.051	1.108
MLP	degraded_big	default	60/60	60/60	215.0	162.5	1.103	1.020	1.026
MLP	degraded_big	worst	60/60	60/60	214.0	162.5	1.102	1.019	1.026

Table 23: E8: composite ledger on the served trajectory. Runs with at least one materialized intervention (incumbent base-gate firings during tenure + adopted meta switches) and mean materialized interventions per run, under the deployment ($K\alpha_{\mathrm{S}}+\alpha_{\mathrm{meta}}$) and rebudgeted ($\alpha_{\mathrm{S}}+\alpha_{\mathrm{meta}}$) conventions.

backbone	scenario	runs w/ materialized interv.		mean interventions (base+meta)	
		deploy	rebudg.	deploy	rebudg.
linear	clean	0/60	0/60	0.00	0.00
linear	degraded	50/60	50/60	1.20	1.40
linear	degraded_both	50/60	50/60	0.83	0.87
mlp	clean	0/60	0/60	0.00	0.00
mlp	degraded	59/60	59/60	1.68	1.80
mlp	degraded_both	58/60	58/60	0.98	0.98

0.40) and at α_{S}/K (rebudgeted, composite level $\alpha_{\mathrm{S}}+\alpha_{\mathrm{meta}}=0.10$); with GCL on top and tenure tracked adoption by adoption. On clean streams the composite certificate's content is measured directly: *zero* runs with any materialized intervention under either convention ($0/60$ per backbone and convention), far inside both composite levels.[3] Under degradation, materialized interventions appear where they should (50–$59/60$ runs adopt), and the theorem's stated price of rebudgeting is visible in the raw: the rebudgeted GCL loss is slightly worse than the deployment-budget one (1.10 vs. 1.08 per step, degraded, linear); validity bought with power: the $\ln K$ shift of (†). One structural finding: in these runs every materialized intervention was a meta adoption; the incumbents' own base gates fired in silico around onset (as E1 measures) but *before* the adopting tenure began, so no base firing landed inside a tenure. The composite ledger covers both kinds; these streams simply exercised one. E8b closes this gap by construction.

E8b: a designed recovery stream that materializes the base half. Because no E8 stream produced an in-tenure base firing, we *designed* one to exercise that branch of Theorem G.8: it is protocol coverage, not a claim about how often such events occur naturally. The scenario degraded-recover applies E1's injected-noise operator to the model's input view (baseline view clean) on a *window* $[t^*, t_{\mathrm{rec}})$ with $t_{\mathrm{rec}} = T_0 + 2(T_1 - T_0)/3$, after which the feed recovers. The intended event order: fallback switches in silico around t^*, GCL adoption of a gated policy, then the *incumbent's* own e-gated recovery switch firing mid-tenure at recovery; matches the raw (Table 24): each run logs the tenure intervals, all base firing times per policy, and every materialized base event with its enclosing tenure, so the overlap is auditable event by event from the raw file alone (the analysis script re-verifies,

[3]The $0/60$ figures are empirical frequencies across independent runs, whereas Theorem G.8 bounds the *probability* of any false materialized intervention on a single trajectory; the frequency is consistent with (but not a direct estimate of) the per-trajectory bound, since each run also spends only part of its budget.

Table 24: E8b: the designed recovery stream. Runs with at least one materialized base firing (the incumbent's own e-gated switch during its tenure), mean materialized base firings and adoptions per run, and median delay from feed recovery to the first in-tenure firing.

backbone	convention	runs w/ mat. base firing	mean base	mean adopted	med. post-recovery delay
linear	deployment	50/60	0.83	1.20	640 h
linear	rebudgeted	40/60	0.67	1.40	1013 h
mlp	deployment	55/60	0.92	1.67	588 h
mlp	rebudgeted	37/60	0.62	1.87	566 h

for every materialized base event, that its time lies inside the recorded tenure and that the firing policy is the incumbent). Both budget conventions are full re-runs, as in E8. The base half of the ledger is now exercised: $50/60$ (linear) and $55/60$ (MLP) deployment-convention runs materialize at least one in-tenure base firing, versus $40/60$ and $37/60$ under the rebudgeted convention, with median recovery-to-firing delays of 640 h vs. 1013 h (linear; the same asymmetry is much smaller for MLP, 588 vs. 566 h; Table 24); the rebudgeted gate fires in fewer runs and later, the $\ln K$ power price a third time. We do not report a false-intervention rate for this scenario: the corruption window makes the base nulls genuinely violated, so these firings are power, not size.

H.14 E10: A REAL EXTERNAL DEPLOYMENT, REPLAYED

Everything so far ran the framework on streams we degraded ourselves and forecasters we fitted ourselves. E10 removes both: the framework consumes a *real* multi-model forecasting deployment; the CDC FluSight forecast hub, where independent teams submitted weekly quantile forecasts of US influenza hospital admissions in actual operation, and none of the forecasts are ours. We take the seven models with sufficient submission coverage over the 2023–24 and 2024–25 seasons (four candidates with real missed weeks or missing jurisdictions were dropped and are listed in the raw; nothing is imputed) plus the hub's official `FluSight-baseline`: the baseline serves only as the base gates' fallback reference and is not in the meta library (no operator would *adopt* the reference model; GCL selects among the seven submitted pipelines), and replay the full stack on the realized panel: $2{,}470$ (week, jurisdiction) cells across 52 weeks and 38–51 jurisdictions per week, with $H = 1$ week of reporting delay (Table 25). It is a *retrospective replay* of a real deployment, not a live intervention; no hospital staffing decision was affected. That is also what makes Assumption G.1 credible here: a replay run after the fact could not have influenced CDC's reporting, so the served trajectory *is* the counterfactual trajectory. The claim holds for the replay design; it is not a verified property of the source system. A *prospective*, pre-registered window is the natural next step; the released code freezes one complete configuration (library rule, budgets, gates, schedules) that any prospective run can execute verbatim. The replayed configuration was not frozen a priori: it was fixed while the replay data existed, and that is not retrospectively verifiable (the constants were carried over from the synthetic-section conventions and rescaled to the weekly panel: γ and the window from E7's per-step values at the panel's step count, the scale grid dyadic around the base window as in E9; the order of operations is not independently verifiable). We do not attempt the prospective window here because a flu season spans October–May and the certificates are only meaningful over a full season, which no review cycle covers. The replay is what we claim; the prospect is made verifiable now: the supplementary file preregisters a full prospective window on the 2026–27 FluSight season (library rule, budgets, gates, schedules, hypotheses, analysis and reporting plan; the E10 configuration verbatim, fixed before that season's first submission), and we anchor it here by its SHA-256 hash: `bbf7e616224724989962e90f827cc4e4ea587b9f928e9887c96dc057385a1cab`. A public timestamp is deferred to camera-ready for anonymity; from the next season on, "frozen a priori" becomes a checkable claim. The prospective window is now *live*: a weekly collection job has been snapshotting the hub's 2026–27 season submissions and truth since August 2026 (pre-season snapshots correctly contain zero submission files, logged as such), so the registered analysis can be executed verbatim on the season's close; the collection log ships with the supplementary package.

Per layer (Figure 8). *Calibration ledger (A1):* the per-model, per-jurisdiction widening recursion moves every pipeline toward the $\alpha = 0.10$ target on the pooled panel; raw submitted 90%-interval miscoverage ranged from 0.120 to 0.381 (the hub baseline: 0.259); calibrated, from 0.077 to 0.166 (0.130), and does not reach it everywhere: the worst model's calibrated miscoverage is still 0.166 at ≈ 50 weekly updates per jurisdiction, the finite-T transient that Theorem 3.1(a)'s $O(1/\gamma T)$ term

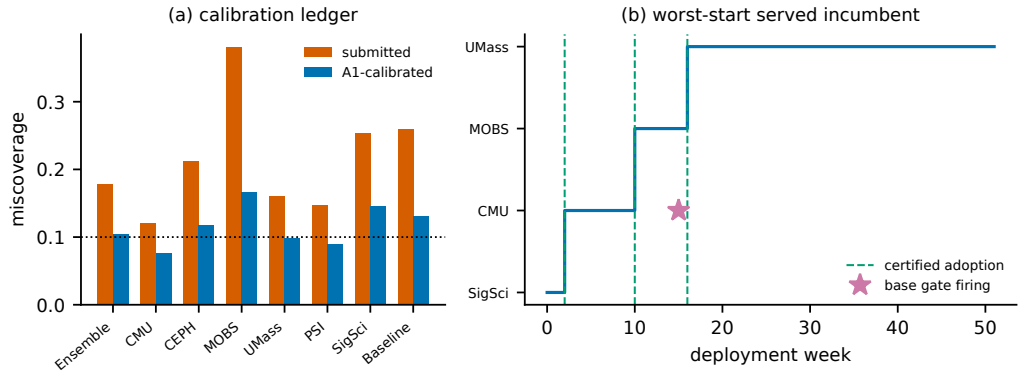

Figure 8: **E10 FluSight replay.** (a) Submitted vs. A1-calibrated 90%-interval miscoverage per pipeline, pooled over the 2,470-cell panel; dotted line = target α. The recursion moves every pipeline toward the budget, not reaching it everywhere at ≈ 50 weekly updates—the finite-T transient of Theorem 3.1(a), on real data. (b) Served incumbent on the *designed* worst-start trajectory (an oracle choice): three certified adoptions (dashed) climb from the ex-post weakest to the best fixed model; the lone base-gate firing (star, CMU) falls outside its tenure, so Theorem G.8 books three materialized interventions against the joint budget.

prices, visible on real data. *Base gates (A3):* six of seven fallback gates (budget α_S/K, monitoring model-vs-hub-baseline) stayed silent over two seasons; correctly, since every model beat the hub baseline on pooled loss, so the gates' nulls are genuinely true and silence is size, not a failure; the seventh (CMU-TimeSeries, the one model with a sustained season-1 stretch behind the baseline) fired once. *Meta layer:* started from the hub's own ensemble (the incumbent an actual operator would serve) GCL adopts *zero* switches in 52 weeks: no competitor's edge over the ensemble (weighted interval score (WIS) per cell, MAD-normalized: 2.664 vs. best fixed 2.643) survives the detectability bar, so the certified answer is "stay," and the served trajectory ties the strongest fixed pipelines. MS-GCL's finest scale adopted one switch (week 27, to a model whose edge on the monitored segment cleared the gate but which is ex post 0.05 worse per cell than the ensemble over the full run, leaving MS-GCL's served trajectory 0.09 per cell behind GCL's certified "stay"): an adoption certified against the *monitored segment*, not against the future, and priced by the $\alpha_{\text{meta}} = 0.05$ false-adoption budget: it is the cost of running more gates on a near-tied library. To give the power half something to detect on real data we also replay from a *designed* worst start (the weakest fixed model, per the E8b convention): GCL then climbs out in three certified adoptions, ending at the best fixed model and cutting served loss from 2.999 (staying put) to 2.704. *Composite ledger:* on the ensemble-start trajectory, zero materialized interventions against a budget of $\alpha_S + \alpha_{\text{meta}} = 0.10$; on the worst-start trajectory, three, all meta adoptions; CMU's lone base firing fell outside its tenure, as the tenure accounting of Theorem G.8 requires. *Decision layer:* the staffing newsvendor of E5 on the calibrated upper endpoints puts the GCL-served pipeline at the best decision loss in the library (13.3 per cell vs. 24.9 for the hub baseline; worst start: 15.1 vs. 21.3 for staying on the weakest model; the MS-GCL served trajectories pay 14.1 and 15.4; their week-27 adoption and slower worst-start exit visible in the decision ledger too). In summary: on a real deployment where the operator's default was already near-optimal, the framework's value was *certified restraint*; calibrated intervals, gates with true nulls left unfired, and a no-switch answer that cost 0.02 WIS per cell against the unknowable best fixed pipeline, and where the start was bad, the certified switches found the exit. T is small at weekly cadence; the contribution at this sample size is anytime validity rather than asymptotics.

H.15 E12: A SECOND EXTERNAL DEPLOYMENT, REPLAYED

To check that E10's findings are not one hub's idiosyncrasy, E12 replays the identical stack: imports the e-process, WIS, newsvendor, and *every* constant from E10; nothing is retuned; on a second real multi-model deployment: the US COVID-19 Forecast Hub, twelve independently operated models submitting weekly quantile forecasts of incident COVID-19 deaths, November 2020 to March 2022 (Table 26). It is an *adjacent* deployment, not an unrelated domain; the same hub infrastructure family, a different disease, target, period, and largely different teams; it tests transfer of the frozen configuration across hubs. The panel keeps 2,610 (week, jurisdiction) cells across 56 of 69 study

Table 25: E10: FluSight replay, pooled over the 2,470-cell panel (weighted interval score (WIS) per cell, normalized by the median absolute deviation (MAD); miscoverage of the submitted vs. A1-calibrated 90% intervals; staffing-newsvendor decision loss per cell). Served rows: GCL / MS-GCL from the hub-ensemble start and from the designed worst start.

pipeline	WIS loss	raw misc.	cal. misc.	decision
hub ensemble	2.664	0.179	0.105	13.3
CMU-TS	2.839	0.120	0.077	17.1
CEPH-Rtrend	2.761	0.212	0.118	15.9
MOBS-GLEAM	2.710	0.381	0.166	17.0
UMass-flusion	2.643	0.160	0.098	13.6
PSI-PROF	2.774	0.147	0.089	16.5
SigSci-TSENS	2.999	0.254	0.145	21.3
hub baseline	3.106	0.259	0.130	24.9
GCL served (ens. start)	2.664	–	–	13.3
MS-GCL served (ens. start)	2.753	–	–	14.1
GCL served (worst start)	2.704	–	–	15.1
MS-GCL served (worst start)	2.841	–	–	15.4

weeks (35–50 jurisdictions per week; weeks with fewer than 20 complete jurisdictions dropped, one of fourteen candidate models dropped at fetch (CEID-Walk, $48/69$ submitted weeks against the 85% rule) nothing imputed), spanning the winter 2020–21 wave, Delta, and Omicron. Per layer: *(A1)* raw submitted miscoverage 0.089–0.600 (hub baseline 0.180) calibrates to 0.047–0.249 (0.103) at ≈ 46 weekly updates per jurisdiction; the same finite-T transient as E10, and the worst pipeline (PSI-DRAFT, raw 0.600) is again not pulled all the way in. *(A3)* nine of twelve fallback gates stayed silent; the three that fired (CU-select, JHUAPL-Bucky, UCSD_NEU-DeepGLEAM) are three of the *four* models behind the hub baseline on pooled loss (2.826, 2.904, 3.284 vs. 2.814); the fourth (PSI-DRAFT, 2.953) trails the baseline but its gate stayed silent: a power miss. No gate fired on a model ahead of the baseline, so size is zero; power caught three of four. *(Meta)* from the hub's own ensemble GCL and MS-GCL adopt *zero* switches in 56 weeks and the served trajectory equals the best fixed pipeline (2.421, which *is* the ensemble); certified restraint replicates on the second hub, and this time MS-GCL's extra gates also produce no firings. From the designed ex-post worst start (UCSD_NEU-DeepGLEAM, 3.284; E10's oracle convention) one adoption is certified at week 4 (MS: week 5) into Karlen-pypm, taking the served loss to 2.521. *(Composite)* zero materialized interventions from the ensemble start; one (the adoption) from the worst start: the worst-start model's own gate firings all land *outside* its tenure, verified event by event in the raw. *(Decision)* E10's newsvendor constants applied verbatim (a resource-provisioning reading, not a claim about any actual staffing decision): served loss 12.7 vs. hub baseline 21.3. The replay caveats of E10 apply verbatim: retrospective, no live decision affected; Assumption G.1 by replay design.

H.16 E13: INDEPENDENT EXTERNAL BASELINES, HEAD-TO-HEAD

All but three comparison policies in E1 are recovered special cases of the framework, so the width separation there could be discounted as by construction. E13 removes that discount: three methods that are *not* instances of the SIP axioms run on the identical E1 protocol; same six streams, same train/calibration/deployment split, same three scenarios, same ten seeds, same target $\alpha = 0.1$, same feedback convention (Table 27).

The baselines. SPCI (Xu & Xie, 2023) predicts the conditional residual quantile with a gradient-boosted quantile regressor on the last 20 residuals, refit every 200 steps on a 600-step window (a faithful but lightened adaptation: the reference implementation refits its regressor more frequently; our refit schedule favors neither side and is documented in the released code header). NexCP (Barber et al., 2023) is weighted split conformal with exponentially decaying weights ($\rho = 0.99$) over the trailing residual history. PH-recal is a non-conformal monitoring baseline in the drift-detection tradition (Page, 1954; Gama et al., 2014): a Page–Hinkley detector on the standardized residual triggers full requantilization on the trailing window on each alarm. All three are recalibration-layer methods on the raw model pipeline: none can repair inputs, fall back, or gate a structural change, which is precisely what the comparison isolates.

Table 26: E12: COVID-19 hub replay, pooled over the 2,610-cell panel (WIS per cell, MAD-normalized; miscoverage of the submitted vs. A1-calibrated 90% intervals; newsvendor decision loss per cell, E10's constants). E10's configuration verbatim.

pipeline	WIS loss	raw misc.	cal. misc.	decision
hub ensemble	2.421	0.159	0.088	12.7
BPagano-RtDriven	2.553	0.275	0.135	16.6
CU-select	2.826	0.444	0.187	23.3
GT-DeepCOVID	2.661	0.201	0.113	15.8
JHUAPL-Bucky	2.904	0.357	0.141	25.5
Karlen-pypm	2.492	0.215	0.110	14.4
MOBS-GLEAM_COVID	2.592	0.318	0.134	14.8
PSI-DRAFT	2.953	0.600	0.249	27.7
RobertWalraven-ESG	2.650	0.383	0.165	18.2
UCSD_NEU-DeepGLEAM	3.284	0.089	0.047	34.9
UMass-MechBayes	2.564	0.133	0.072	15.3
USC-SI_kJalpha	2.594	0.557	0.233	16.4
hub baseline	2.814	0.180	0.103	21.3
GCL served (ens. start)	2.421	–	–	12.7
MS-GCL served (ens. start)	2.421	–	–	12.7
GCL served (worst start)	2.521	–	–	14.8
MS-GCL served (worst start)	2.525	–	–	14.9

Table 27: E13: external baselines vs. the unified policy, E1 protocol (linear backbone; MLP agrees; pooled over six streams; mean $\pm$ sd over per-stream seed-means). *shortfall* = max over runs of $(0.90 - \text{coverage})_+$; external rows repeat under *degraded-both* (they never touch the repair feed); SPCI is a lightened adaptation (refit every 200 steps).

policy	cov.	width	shortfall
clean			
SPCI	0.881	0.24 ± 0.12	0.027
NexCP	0.910	0.28 ± 0.14	0.000
PH-recal	0.923	0.31 ± 0.13	0.012
ACI (R1)	0.900	0.27 ± 0.13	0.001
Unified DRC	0.900	0.27 ± 0.13	0.001
degraded			
SPCI	0.872	1.61 ± 0.49	0.036
NexCP	0.903	1.75 ± 0.52	0.002
PH-recal	0.898	1.68 ± 0.50	0.032
ACI (R1)	0.884	1.65 ± 0.48	0.022
Unified DRC	0.900	0.65 ± 0.22	0.006
degraded-both			
SPCI	0.872	1.61 ± 0.49	0.036
NexCP	0.903	1.75 ± 0.52	0.002
PH-recal	0.898	1.68 ± 0.50	0.032
ACI (R1)	0.884	1.65 ± 0.48	0.022
Unified DRC	0.887	1.49 ± 0.41	0.020

Findings (Table 27; linear backbone shown, MLP parallel). On clean streams the unified policy is indistinguishable from ACI (width 0.274 vs 0.271, coverage 0.900); SPCI undercovers (0.881, max shortfall 0.027), NexCP holds the budget at 2% wider intervals, PH-recal overcovers at 11% wider. Under degraded, the separation is structural: the unified policy repairs the corrupted view and holds coverage 0.900 at width 0.646; every external baseline, unable to act on the pipeline, widens to 1.61–1.75; 2.5–2.7$\times$ (2.2–2.4$\times$ MLP; all width differences significant at the stream level, Holm-adjusted cluster bootstrap). NexCP is the one external that still holds the budget there (max shortfall 0.002; its decayed-weight quantile tracks the shift, at full width price); SPCI (0.036) and PH-recal (0.032) breach it, as does ACI (0.022). Under degraded-both the external rows are unchanged (they never touch the repair feed, so the second corruption is invisible to them) while the unified policy loses its instrument: width 1.49, coverage 0.887 (max shortfall 0.020), i.e. a $\leq 12\%$ residual width edge over SPCI/PH-recal and *worse* budget compliance than NexCP. The advantage under degraded is the value of actuators when a repairable view exists, not a generic dominance claim; where no repair is possible, a well-tuned recalibration-only method is the right tool.

Table 28: E14 budget-sensitivity scan: the identical protocol re-run at $\alpha \in \{0.05, 0.20\}$, the robustness budget halved and doubled relative to the base $\alpha = 0.10$. *viol. rate* = realized certificate-violation frequency of the direct (R9) ledger; *profit ratio* = its mean realized arbitrage profit over the better of the two coverage-route baselines on the same stream. † marks the two cells missing the registered pinning criterion $|\text{viol.} - \alpha| \leq 0.002$, both in the conservative (undershoot) direction.

α	metric	NP	PJM	BE	DE	FR
0.05	viol. rate	0.0498	0.0498	0.0499	0.0495	0.0500
	profit ratio	2.33×	2.09×	2.28×	2.83×	2.62×
0.20	viol. rate	0.1970†	0.1945†	0.1999	0.1996	0.1995
	profit ratio	1.55×	1.56×	1.26×	1.77×	1.29×

H.17 E14: A CRO DECISION TASK WITH EXTERNALLY DEFINED LOSS

E5's newsvendor and storage costs are parameters we chose; a skeptical reviewer asks for a decision task whose loss is not ours. E14 (protocol registered before execution, hash anchored in the released preregistration file) adopts the CRO interface verbatim from the external literature (Hu et al., 2026; Sun et al., 2023): single-period storage arbitrage; charge fraction $z_t \in [0, 1]$ at the known price p_t, sell at the unknown next-hour price y_t, loss $\phi(y, z) = z\,(p_t - y)$; with the minimax map $z_t = \arg\min_z \max_{y \in U_t} \phi(y, z)$, served certificate $r_t = \max_{y \in U_t} \phi(y, z_t)$, and robustness functional $\mathbb{P}\{\phi(y_t, z_t) \leq r_t\} \geq 1 - \alpha$ at $\alpha = 0.10$, on the five public EPF price streams of Appendix H.1's data package (51,415 scored hours each; 24-hour-lag persistence forecaster with trailing-MAD scale, fixed in the preregistration; deterministic, no seeds).

Policies. static CRO (split-conformal coverage-route CRO, the conventional baseline of Hu et al., 2026: radius frozen after a 1,000-hour calibration burn-in); CRO-ACI (coverage-route with the A1 recursion on the coverage indicator); R9 (the R9 functional of Proposition F.6: the A1 recursion runs directly on the decision-level violation indicator $\mathbf{1}\{\phi(y_t, z_t) > r_t\}$). CRC-Robust's own batch set-construction optimizer is not reimplemented (no public code at the time of writing); E14 tests their *constraint*, not their algorithm.

Findings. The R9 ledger pins the realized violation frequency at 0.0997–0.1002 on all five streams: Theorem 3.1(a) on the transplanted functional, H1 of the registered protocol confirmed. Both coverage-route policies over-satisfy the robustness target everywhere (realized violations 0.016–0.045 against the allowed 0.10): coverage is a sufficient but not necessary condition, measured sequentially; the sequential analogue of CRC-Robust's central observation, and H2 confirmed: at the same target level, the direct ledger's mean realized arbitrage profit per hour is 1.85–2.45× CRO-ACI's on every stream, and 1.79–2.45× the better of the two coverage baselines per stream (e.g. NP: 0.151 vs 0.067/0.074 MAD-normalized units; PJM: 0.549 vs 0.241/0.298; all ratios recomputable from the released raw outputs). Two self-positioning remarks. First, H1 is Theorem 3.1(a) executing, not an empirical discovery: the A1 recursion runs directly on the violation indicator, so the pinned frequency is the mechanical content of the theorem, and what H1 confirms is implementation correctness; the empirical content of E14 is H2 (and H3's outcome below). Second, part of the profit gap flows through a trading-frequency channel: with the linear loss the interval minimax decision is bang-bang ($z_t \in \{0, 1\}$, charge iff the certified lower price bound exceeds the current price), so the direct ledger's tighter certificates clear the entry condition more often and trade more hours; this does not change the direction of H2 (same target level, more realized profit) but the gap should not be read as a pure per-trade improvement. One registered hypothesis *failed*: H3 predicted static CRO would drift *above* the violation target on at least one stream; in fact it drifted conservative on all five (violations 0.016–0.045), so on these streams the static method's miscalibration showed up as lost efficiency; validity held. The delimitation stands: E14 exercises the sequential trajectory axis (ledgers under recalibration), not the set-construction-optimization axis that is Hu et al. (2026)'s contribution.

Sensitivity to the budget level (preregistered; Table 28). A scan re-runs the identical protocol—same streams, forecaster, policies, and every other scalar—at $\alpha \in \{0.05, 0.20\}$, i.e. the robustness budget halved and doubled. The profit direction never reverses: on all ten (α, stream) cells the direct ledger's realized arbitrage profit exceeds the better of the two coverage-route baselines, by 1.26–2.83× (at $\alpha = 0.05$: 2.09–2.83×; at $\alpha = 0.20$: 1.26–1.77×; base $\alpha = 0.10$: 1.79–2.45×). The registered pinning criterion $|\text{viol.} - \alpha| \leq 0.002$ holds in eight of ten cells; the two misses, NP (0.1970) and PJM (0.1945) at $\alpha = 0.20$, are both in the conservative direction (undershoot; the budget is never

Table 29: Experiment-to-artifact map of the released package. Scripts reside in `code/`, raw outputs in `raw/`; wildcard patterns abbreviate per-backbone suffixes (`_mlp`, `_tsfm`, `_tsfm2`).

Content	Run scripts	Raw outputs
E1–E4 core suite (all backbones)	`unified_exp*`, `natural_e*` `e1_acibase/`, `e1_h24_sens/`	`results_e{1,3,4}*`, `raw_e1_*`
E5/E5c decision layer	`decision_e5*`, `sens_e5c`	`results_e5*`, `tests_e5*`
E2 trajectories	`traj_e2`	`traj_e2`
E6–E9, E11 meta/composite	`meta_e*`, `composite_e8*`	`results_e{6--9,11}*`
E10/E12 hub replays	`fetch_*`, `e10`, `e12`	`results_e{10,12}`
E13 external baselines	`external_e13`	`results_e13*`
E14/E15 preregistered subpackages	`e14_*/`, `e15_*/`	`raw_e14`, `raw_e15`, `raw_e14_sens`
Aggregates, tests, figures	`analyze_e*`, `gen_table*`, `make_figures*`	`summary_*`, `tests_*`

overspent). A post-hoc diagnostic, outside the registered protocol and not used to alter any judgment, attributes the two undershoots to the $q \geq 0$ lower boundary of the A1 recursion binding at the loose budget on those two streams (clipped on $1.69\%/2.72\%$ of steps, with $z_t = 0$ steps—structurally unable to violate—at $73.8\%/72.1\%$); $\alpha = 0.05$ and the base $\alpha = 0.10$ sit far from that boundary and show no such effect.

H.18 E15: BEYOND FORECASTING; ONLINE CLASSIFICATION RISK CONTROL

Every stream above is a forecasting stream; the framework's claim is not about forecasting. E15 (protocol registered before execution, hash anchored in the released preregistration file) instantiates the deterministic ledger on the served-threshold filter interface of Section 1 with no forecasting component: online multiclass set-valued classification on UCI Covertype ($581{,}012$ rows, 7 classes) consumed in its *native row order*; spatially structured survey order, a natural source of covariate and label drift with no time index. A multinomial logistic model is trained on the first $50{,}000$ rows and frozen; the served output at each of the $521{,}012$ scored steps is the score-threshold set $U_t = \{k : p_k(x_t) \geq 1 - q_t\}$, and the booked proxy is $\mathbf{1}\{y_t \notin U_t\}$ at $\alpha = 0.10$. Two policies share the model and interface: a split-conformal threshold frozen after a $10{,}000$-row calibration block, and the A1 recursion ($\gamma = 0.01$, fixed in the registered protocol) initialized at the same threshold; the SIP with an empty structural library.

Findings. The static threshold drifts far off budget on the naturally ordered stream: realized miscoverage 0.371 against the allowed 0.10, climbing steadily across the stream (segment path $0.170 \to 0.371$); H2 confirmed, and the same failure mode the forecasting experiments exhibit, on a task with no forecasting in it. The A1 recursion holds the booked proxy at 0.1011 over half a million steps: H1, the theorem executing on a new domain. One registered guardrail *failed*: H3 predicted the ledger's mean served set size would stay within $1.5\times$ the static baseline's, and it did not (1.98 vs 1.12 classes, a $1.77\times$ ratio), with the caveat that the two sizes are not at comparable risk (the static baseline's smaller sets are bought at $3.7\times$ the target violation rate). E15 exercises the deterministic ledger only; the structural library and the decision layer are not instantiated here, so this is domain evidence for the accounting core, not for the full framework.

I EXPERIMENTAL DETAILS AND REPRODUCIBILITY

Table 29 maps each experiment to the released artifacts that produce it; run scripts live under `code/`, raw outputs and aggregates under `raw/`, and stream data under `data/` in the released package. The added calibration baselines, the feedback-delay sensitivity run, E14, and E15 ship as self-contained subpackages (script, and for E14/E15 the frozen preregistration, with raw output together).

Streams. EPF markets NP, PJM, BE, FR, DE: hourly day-ahead prices from the open benchmark of Lago et al. (2021), truncated to the first 20,000 points. ETTh1: hourly oil temperature (OT column) (Zhou et al., 2021), first 17,000 points. Splits: train 50% / calibration 15% / deployment 35%. All six stream CSVs and their loaders are included in the released package, so all four backbones' E1 (and E2) re-run end-to-end from the package alone, like E3 and E4.

Pipelines. Model (main backbone): linear AR on lags $(1, 2, 3, 24, 25, 168)$ with intercept, least squares on the training split, frozen. Model (second backbone; Appendix H.8): shallow MLP on the same lag features; two hidden layers $(64, 32)$, ReLU, Adam, max 300 iterations with early stopping, one seeded fit per run (`scikit-learn` `MLPRegressor`); trained on the same split and frozen. Models (third and fourth backbones; Appendix H.9): Chronos-Bolt small (Ansari et al., 2024) and TimesFM-2.5 200M (Das et al., 2024), both zero-shot (no fitting to the streams): median forecast (Chronos: 0.5-quantile head; TimesFM: continuous-quantile head, median column) from 512-step contexts of the model's (possibly degraded) input view, non-overlapping 24-step forecast blocks, float32 on one GPU; the first 512 steps of each stream carry no TSFM forecast and are excluded from deployment scoring for those backbones. The four backbones are run as separate full replications of the design and released as separate raw files; they are never pooled. Baseline: seasonal naive with period 168 (EPF) or 24 (ETTh1). Repair: trailing rolling median (window 5) applied to the model's input view, computable from data available at each step. Degradation: additive Gaussian noise, sd $= 1.5\times$ stream sd, applied to the model's input view from onset t^* (one third into deployment); the baseline's view is unaffected (upstream feature corruption).

Budget and gates. $\alpha = 0.10$, $\alpha_{\mathrm{S}} = 0.05$, recursion step $\gamma = 0.02$ (in units of calibration sd), e-process: 20-point geometric λ-grid with $B_D = 4$, window $W = 500$; the fallback and recovery gates each receive budget $\alpha_{\mathrm{S}}/2$ with per-segment spending $\alpha_j = (\alpha_{\mathrm{S}}/2)/j(j+1)$, so the joint spend over both gates is $\leq \alpha_{\mathrm{S}}$, matching Axiom 2.5 exactly. Conformal PI integrator: $0.5\,\sigma\tan(\mathrm{clip}(0.05\sum_s(\mathrm{err}_s - \alpha), \pm 1.4))$ (saturated, non-negligible gain). Debt trigger: exponentially weighted coverage (decay 0.98) below $1 - \alpha - 0.05$, cooldown 24 steps; costs $c(\text{repair}) = 1$, $c(\text{switch}) = 0.2$ per fallback step. Added calibration baselines (calibration-only; no repair, gating, or structural access): `aci_dt` is DtACI (Gibbs & Candès, 2024) ported to the radius representation — $K = 8$ experts at geometric step sizes $\gamma_k \in \{2^{-3}, \ldots, 2^4\}\gamma$, each maintaining its own additive offset recursion; expert weights updated by exponential weights on the level-$(1-\alpha)$ pinball loss ($\eta = 2.72$) with fixed-share smoothing ($\sigma = 10^{-3}$, the authors' defaults); the served radius uses one expert's offset sampled from the current weights, seeded deterministically per (stream, scenario, seed) for reproducibility. `aci_bet` is Algorithm 1 of Podkopaev et al. (2024) verbatim on σ-normalized scores ($W_0 = 1$, $\lambda_1 = 0$, $s_1 = 0$; $g_t = \alpha - \mathrm{err}_t$, $W_t = W_{t-1} - g_t s_t$, $\lambda_{t+1} = \frac{t}{t+1}\lambda_t - \frac{g_t}{t+1}$, $s_{t+1} = \lambda_{t+1} W_t$), the served radius $\max(\varepsilon, s_t)$ clipped only at the same feasibility floor every policy uses; the unconstrained update is kept in state. Feedback delay $H = 1$. Sensitivity: re-running E1 (linear) with the loss of step t booked only at $t + 24$ leaves the picture unchanged—under degradation the unified policy holds coverage 0.900 at width 0.71 against conformal PI's 1.88 ($2.6\times$, same ordering as $H{=}1$); the delay's price is a modest uniform width inflation and a higher clean repair overhead (65.7 vs. 43.5 cost units).

Natural streams (E3). ERCOT hourly system load, two windows: *Uri* (2020-09-13 to 2021-07-31, 7,712 h; event window hours $[3680, 3824)$ = 2021-02-14 to 2021-02-20, Winter Storm Uri) and *COVID* (2019-09-14 to 2020-08-31, 7,712 h; event window hours $[4472, 5504) \approx$ 2020-03-19 to 2020-05-01). Event windows are fixed from public incident records (FERC–NERC Uri report; spring-2020 lockdown orders) and shipped in the raw metadata; nothing is injected or modified. Deployment starts 720 h before event onset; calibration is the last 15% of the pre-deployment data, AR training the rest; baseline period 168; repair operator identical to E1. One deterministic run per (stream, policy); raw outputs are released. Both ERCOT CSVs ship in the package, so E3 re-runs end-to-end from it alone.

Asymmetric incident streams (E4). Two public benchmarks (Autoformer/Time-Series-Library): PEMS road occupancy (862 sensors, hourly, 2016-07-01 to 2018-07-02, 17,544 h) and ECL client load (321 clients, hourly, 2016-07-01 to 2019-07-02, 26,304 h). Incident rule (uniform over both files): longest maximal zero-run ≥ 100 h starting after hour 2000, history before onset $< 2\%$ zeros, and an eligible target exists; this yields feed sensor 181 (terminal outage from hour 16,225; 1.5% zeros before) and feed client 182 (terminal outage from hour 6,390; 0.02% zeros before), and excludes electricity client 245 (qualifying zero-run, no target with pre-failure correlation ≥ 0.85). Target sensors: deterministic selection on pre-deployment data (healthy sensors, $< 1\%$ zeros, pre-failure correlation ≥ 0.85 with the feed; minimize nowcast-vs-baseline mean absolute-loss gap on the calibration split), yielding sensor 502 and client 253. Model: linear nowcast on feed lags $(1, 2, 3, 24, 25, 168)$, frozen; baseline: seasonal naive (period 168) on the target's own history; repair: weekly-seasonal imputation of dead feed readings. Deployment starts 720 h before onset; splits, budget, and gates identical to E3. Raw outputs (including gate firing times) are released; the package

provides verified download scripts for both benchmarks plus per-incident extracts (date, feed, target columns) that replay E4 without the ∼230 MB downloads.

TSFM natural replays (E3/E4). The E3 replay is identical to the linear one with the model pipeline replaced by zero-shot Chronos-Bolt on the (repaired or raw) load series. For E4, both TSFMs are univariate and cannot condition on the feed as a covariate; the model pipeline is therefore a *virtual sensor*: an affine feed→target mapping fit on the training split only, forecast by the TSFM from the mapped feed's context, so the feed outage still degrades only the model view. This design choice is a limitation (Section 6): it tests the intervention loop around a TSFM, not covariate-conditioned TSFM forecasting. Raw outputs are released separately for Chronos-Bolt and TimesFM-2.5.

Decision layer (E5). Served forecasts and widths are re-generated per (policy, stream, scenario, seed) by the same protocol as E1 (equivalence with the E1 implementation is asserted programmatically at run time), then routed through the two interfaces over the *full deployment segment* (the pre-onset prefix, roughly the first third, included); degraded-scenario aggregates therefore mix in clean-prefix steps, which dilutes rather than inflates the reported separations. Newsvendor: $h = 1$, $b = \rho/(1-\rho) \cdot h = 15.6$ from $\rho = m/(m + r\ell/52)$ with margin $m = 0.30$, holding rate $r = 0.25$/yr, cycle $\ell = 4$ weeks; order $z_t = L_t + \rho(U_t - L_t)$; all quantities in calibration-sd units. Storage: unit energy capacity, round-trip efficiency $\eta = 0.9$, one buy hour b and one sell hour $s > b$ per non-overlapping 24-h block, chosen to maximize the certified floor $\pi(b, s) = \eta L_s - U_b$; realized profit $\eta y_s - y_b$; the five EPF price streams only (a storage device arbitrages prices, not temperatures). Raw outputs and the Wilcoxon/Holm test statistics are released per backbone. Table 16 reports the MLP replication; Tables 17 and 18 the TSFM decision replications; Tables 10, 11 and 13 the TSFM natural replays. E5c (block ledger; Appendix H.10): same storage protocol with the inflated bands $[L_t - \lambda_d, U_t + \lambda_d]$ and recursion Equation (2) ($\gamma_D = 0.05$, $\alpha_D = 0.1$, $\lambda_1 = 0$); the uncontrolled ($\lambda \equiv 0$) schedule and floor are logged from the same days for the paired comparison; raw outputs and aggregates are released.

Composite ledger (E8/E8b). E8 re-runs the seven intervention-library policies per configuration at both budget conventions and runs GCL (α_{meta}, W as E7) on the resulting loss matrices; E8b's degraded-recover scenario replaces the terminal noise window by $[t^*, t_{\text{rec}})$ as described in Appendix H.13. For exact replayability we note that in degraded-recover the repair view applies the rolling-median smoothing over the *whole* deployment segment, including the post-recovery clean suffix: the same operator and extent as every other scenario, so repair costs on the clean suffix are real, as on clean streams. E9 runs MS-GCL (grid $\{125, 250, 500, 1000\}$, per-scale budget $\alpha_{\text{meta}}/4$) and single-scale GCL on identical loss matrices; its degraded-big scenario is E1's degraded with noise strength 6.0 instead of 1.5 (model view only; repair smoothing applied over the whole deployment segment, as everywhere). E11 runs SR-GCL: MS-GCL with per-scale candidate locking: scale i re-nominates only at the boundaries of its own W_i-step matured windows, a candidate change at a boundary advances that scale's segment counter and resets its gate, and between boundaries the locked candidate's confirmation runs uninterrupted; against MS-GCL and single-scale GCL on identical loss matrices, from the default start and from the ex-post worst post-onset policy (pooled matured loss argmax over the post-onset segment; oracle construction, described in Appendix H.12).

FluSight replay (E10). Snapshot fetched from the public `cdcepi/FluSight-forecast-hub` repository by the released fetch script (snapshot timestamp recorded in the released metadata and in the raw outputs): 1-week-ahead (`horizon 1`) quantile submissions for `wk inc flu hosp`, quantile levels $\{0.05, 0.25, 0.5, 0.75, 0.95\}$, all reported jurisdictions, Saturdays 2023-10-14 to 2024-05-04 and 2024-11-23 to 2025-05-31; truth from the hub's released admissions target series. Model inclusion is mechanical: of twelve candidate models, those with $\geq 85\%$ submitted weeks are kept by the fetch (`LosAlamos-DoSiDo`, `GT-FluFNP`, `ISU_NiemiLab-ENS` fail), and models with $< 80\%$ of the baseline's (week, jurisdiction) cell coverage are dropped by the released replay script (`NIH-Flu_ARIMA`: only 13 jurisdictions submitted in season 1); the panel keeps (week, jurisdiction) cells where all seven kept non-baseline models, the hub baseline, and the truth are present (weeks with < 20 such jurisdictions dropped; per-week counts in the raw). Losses: WIS restricted to the five snapshot quantiles, normalized per cell by the trailing 4-week MAD of that jurisdiction's truth (floor 1), clipped at $B = 4$. Calibration: per (model, jurisdiction) additive widening of $[q_{0.05}, q_{0.95}]$ in units of the submitted interval half-width, $\gamma = 0.5$, $\lambda \in [0, 10]$. Gates: α_{S}/K fallback gates per model vs. the hub baseline and GCL/MS-GCL at α_{meta} (split across scales $\{100, 250, 500\}$ panel steps for MS; the baseline is the gates' fallback reference only and is excluded from the meta library), all on matured losses with $H = 1$ week (a week's cells mature together when the next week begins;

adoptions are weekly). Filtration convention for the block maturation: gates step through a matured week's cells in fixed jurisdiction order, so the e-process null is per panel step, conditional on all previous weeks *and* the preceding cells of the same week; the per-cell conditional-mean nulls of Theorems 3.1 and G.3 are exactly of this form, so Ville's inequality applies unchanged; only the *decision* cadence (adoptions, firings applied from the next week) is weekly. The "worst start" of the designed variant is selected *ex post* as the argmax of pooled full-run WIS among the seven kept non-baseline models (`SigSci-TSENS`): an oracle choice by construction, consistent with its role as protocol coverage (giving the power half a detectable gap), not a deployable rule. MS-GCL served decision losses in Table 25 are computed from the same snapshot and recursion as the GCL rows. Newsvendor costs as E5 ($h = 1$, $b = 15.6$), order = calibrated upper endpoint, loss MAD-normalized per cell. The replay script is self-contained given the snapshot (the e-process and cost constants are inlined, byte-identical to the released harness's implementations). Raw outputs and aggregates are released.

COVID-19 hub replay (E12). Snapshot fetched from the public `reichlab/covid19-forecast-hub` repository by the released fetch script (snapshot timestamp recorded in the released metadata and in the raw outputs): `1 wk ahead inc death` quantile submissions at levels $\{0.05, 0.25, 0.5, 0.75, 0.95\}$ (parsed numerically; submission files vary in formatting), all reported locations, study weeks ending 2020-11-21 to 2022-03-12; for each study week the fetch tries the Sunday and Monday submission dates. Truth: the hub's daily JHU incident-death series (git-LFS), aggregated to weekly sums per state/national location keyed by the Saturday week-ending date; the hub's weekly truth convention. Model inclusion is mechanical: of fourteen candidates, models with $\geq 85\%$ submitted study weeks are kept at fetch (`CEID-Walk` fails at $48/69$); the panel keeps (week, location) cells where all twelve kept non-baseline models, the hub baseline, and the truth are present, and weeks with fewer than 20 such locations are dropped (56 of 69 weeks survive; per-week counts in the raw). One submission file contains literal `NULL` values; the affected (model, week, location, quantile) rows are skipped, and cells missing any of the five quantiles are excluded like any other incomplete cell: never imputed. Everything downstream is imported unchanged from the FluSight replay: losses (five-quantile WIS, trailing 4-week MAD normalization with floor 1, clip $B = 4$), calibration ($\gamma = 0.5$, $\lambda \in [0, 10]$), gates (α_{S}/K per model vs. `COVIDhub-baseline`; GCL at $W = 250$ and MS-GCL over $\{100, 250, 500\}$ at α_{meta}, baseline excluded from the meta library), block-maturation filtration, $H = 1$ week, and the E5 newsvendor constants. Meta starts: `COVIDhub-ensemble` (the hub's served aggregate) and the ex-post pooled-WIS argmax (`UCSD_NEU-DeepGLEAM`), the oracle convention of E10. Raw outputs (including per-week incumbents and all base-firing times, so tenure containment is auditable from the raw alone) and aggregates are released.

External baselines (E13). The external-baseline driver reuses the suite's stream builder and policy runner verbatim, so streams, splits, scenarios, seeds, and the unified/ACI reference rows are bit-identical to the E1 protocol. SPCI: gradient-boosted quantile regressor (`scikit-learn`, 25 trees, depth 3, quantile loss at $1 - \alpha$) on the last $p = 20$ absolute residuals, refit every 200 steps on a 600-residual window, seeded per run; the served interval is the model forecast $\pm$ the predicted residual quantile. NexCP: weighted empirical quantile of trailing absolute residuals (history cap 2,000) with weights ρ^k, $\rho = 0.99$, plus the point mass at $+\infty$. PH-recal: Page–Hinkley on the standardized absolute residual ($\delta = 0.005$, threshold $\lambda = 8$); on alarm, requantilize to the empirical $(1 - \alpha)$-quantile of the last 300 residuals and reset the detector. Constants were fixed before any result was inspected and are not tuned per stream. Raw outputs and the Wilcoxon/Holm width tests against the unified policy are released.

Statistics. Paired tests at the stream level: the pairing unit is the stream ($n = 6$), with per-stream seed-means—seeds are replicates within one underlying stream, not independent units. Exact Wilcoxon signed-rank (Wilcoxon, 1945) (granularity floor $p = 0.031$ at n=6) plus a paired cluster bootstrap over streams, both Holm-corrected (Holm, 1979) across the eight comparisons per (scenario, metric); the bootstrap is the powered statistic. Per-stream direction consistency is reported alongside. Run-level and config-level aggregates and all test statistics are released per backbone.

Compute. The linear and MLP suites and the decision layer run on CPU (under one hour, about two hours, and under one hour respectively); each TSFM replication runs zero-shot inference for $\sim$180 full-stream forecast configurations on a single consumer GPU (about six hours for Chronos-Bolt; TimesFM-2.5 is served from a content-hash forecast cache that deduplicates identical input views, about five hours).

Per-stream results. Table 3 reports coverage per stream for the degraded scenario. Table 5 reports the degraded-both scenario (Appendix H.4) and Table 2 the violation-threshold sensitivity analysis.

J Full Limitations Discussion

This is the unabridged limitations discussion summarized in Section 6.

(i) Theorem 3.1(a) is a long-run realized-frequency guarantee; it does not provide finite-sample distributional coverage, and per-instance conditional control is impossible without assumptions (Foygel Barber et al., 2021). (ii) Theorem C.1 assumes a persistent effect size Δ; slowly ramping degradations weaken the delay bound. (iii) The experiments use four backbone families but only two structural actuators; the theory covers richer libraries (abstention, operator subsets) that we exercise only through the recovered special cases. (iv) The clean-stream repair overhead shows that ungated triggers inside the unified policy remain a cost center; gating the repair trigger is possible in the framework but consumes α_{S} budget, a trade-off we leave to future work. (v) The TSFM evidence is two models from two families (Chronos-Bolt small and TimesFM-2.5), and their E4 incidents use a virtual-sensor mapping because both models are univariate; we claim only what Appendix H.9 shows under these backbones (frontier vanishing under injected noise, reappearing on structural incidents), not foundation-model generality. (vi) The decision layer's per-step storage floors are built from marginally calibrated bands and carry no day-level guarantee by themselves; the block ledger of Theorem F.10 controls the long-run day-level violation *frequency* (measured in E5c), but per-day or distributional day-level claims remain out of reach, as per-instance conditional coverage does at step level. (vii) Theorem G.2's regret is to the best fixed library member and its switches are ungated; Theorem G.3 certifies adopted switches and tracks switching comparators, but only on the (Δ, W)-detectable class; below the detection margin, only the certificate survives, and Proposition G.4 shows some such price is intrinsic (Corollary G.5 matches the lower bound when the served gap is of order Δ, and Theorem G.6 closes the worst-case B/Δ factor on the gap-homogeneous detectable class; Theorem G.7 on the c-level-stable heterogeneous class; only unstable gap profiles remain open); Theorem G.8 closes the served-trajectory ledger under the joint budget $\alpha_{\mathrm{S}} + \alpha_{\mathrm{meta}}$, but the in-silico ledgers of non-incumbent policies remain simulation statements, and ungated repairs stay outside the composite certificate. The meta theorems require no performativity (Assumption G.1), which holds in our experiments but not in feedback-coupled deployments; Proposition G.10 prices the counterfactual readings under ε-performativity ($+k_{\mathrm{ref}} E_T$ on every probability level) rather than removing the assumption; under persistent coupling the certificates are vacuous. (viii) E10 and E12 extend the evidence beyond streams we constructed: two real multi-team epidemic-forecasting deployments consumed end to end, the second consuming the first's configuration verbatim, but they are *retrospective replays*: no live decision was affected; Assumption G.1 holds by the replay design rather than by verification of the source system, and ≈ 50 weekly steps per jurisdiction mean the anytime ledgers, not the asymptotic regret terms, carry the content there; the two hubs share an infrastructure family, so E12 is evidence of cross-hub transfer, not of domain generality. All other experimental streams are time series we degraded or replayed ourselves; the SIP formalism is stated for general sequential prediction systems.