TRUST-GATED, MODEL-ASSISTED SUPERVISION OF DEPLOYMENT-TIME ADAPTIVE FILTERS: A BENCHMARK AND AN EMPIRICAL ACCOUNT OF SUPERVISION TIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive filters run on billions of devices, and nearly all of them throttle their adaptation rate. Aggressive adaptation diverges when the plant drifts from its calibration, and in active noise control (ANC), our setting, one screech into a user's ear outweighs decibels of average-case gain. Learned policies could drive adaptation harder, but nothing bounds a misgeneralizing policy in a live signal path, and model-based safe-RL certificates need the plant knowledge that goes stale. We formalize safe deployment-time adaptation, contribute a GPU-vectorized headphone-ANC benchmark with ear re-seats and calibration faults, and measure what supervision costs under matched controls (same in-loop checkpoint, only the supervisor varies, plus an on-policy protocol trained natively per tier) against a hand-tuned watchdog, CUSUM and EWMA null-calibrated offline at equal false-alarm budget. Our supervisor, *Evidence-Gated Adaptation* (EGA), has two tiers. The trust-gated reflex tier (randomized anchor probes, certificate-free rollbacks, and a virtual-probe channel validated by an anytime-valid e-process) trails the on-policy watchdog by a small paired deficit on the main suite (-0.14 dB, CI $[-0.19, -0.11]$) but beats it under a polarity miscalibration fault ($+0.70$ dB, CI $[+0.60, +0.81]$). Under a duty-cycled harm adversary it ends 12.3% of episodes in severe harm, against 21.4/21.2% for budget-matched CUSUM/EWMA charts fed the same virtual evidence and 38.3% for the watchdog. The optional certified tier (mixture e-processes, trial-certified commits, demotion, and a logged harm ledger) costs a further $+0.94$ dB paired. It adds guarantees and an audit trail without lowering any harm rate. Benchmark and code are provided as anonymized supplementary material.

1 INTRODUCTION

Adaptive filters cancel noise in headphones, echo in calls, and interference in modems (Widrow & Stearns, 1985; Kuo & Morgan, 1996). LMS and its filtered-x variant (FxLMS) for active noise control (Morgan, 1980; Elliott & Nelson, 1993) adapt a linear filter by stochastic gradient descent with step size μ: large steps track nonstationarity but destabilize catastrophically when the secondary-path estimate's phase is wrong (a real ear after a re-seat), while small steps forfeit most of the attenuation. Production systems therefore use conservative, hand-tuned schedules.

Learned adaptation for ANC exists: offline filter selection (Luo et al., 2024; Shi et al., 2022), learned update rules (Casebeer et al., 2023; Haubner et al., 2024), RL-tuned step sizes (Pepin et al., 2021, preprint). None of it carries a guarantee. A misgeneralizing policy can amplify noise into the ear indefinitely, and to our knowledge none runs in a deployed device's live signal path. Classical safe RL does not close the gap, because certificate-based methods assume a known plant model (Berkenkamp et al., 2017; Chow et al., 2018; Cheng et al., 2019; Achiam et al., 2017), and here the certificate-relevant quantity, the true secondary path S, is what drifts from its factory calibration $\hat{S}$.

We rely instead on two properties that adaptive filters and many parametric controllers have, and that robots do not. Because filter parameters can be hot-swapped instantly, a known-safe fallback

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.

costs one assignment. Error energy is measured continuously in an environment that does not react to the controller, which makes paired comparisons moments apart meaningful. The supervisor therefore needs no plant model. It monitors adaptation, retreats cheaply and often, and asks for positive evidence only before a promotion. *Evidence-Gated Adaptation* (EGA) implements this in two tiers. Certificate-free reflexes with trust-gated model assistance do the deployment work, and formal e-process certification is an optional tier whose cost we measure under matched controls (Sec. 6). Certification adds auditability, guarantees and a tunable budget. It does not improve harm rates or benign decibels, both of which come from the reflex tier.

Contributions. *(1) Problem class + benchmark.* We formalize *safe deployment-time adaptation*: an arbitrary adaptation process proposes filter updates. A supervisor with hot-swap, rollback and a bounded-output guard must bound realized harm against a trusted fallback with no plant model (Sec. 3), instantiated in a GPU-vectorized ANC benchmark with randomized acoustics, re-seats that break the FxLMS phase condition, and calibration-fault suites (Sec. 6). *(2) A trust-gated reflex supervisor.* EGA's deployment tier (Sec. 4) uses randomized anchor probes and certificate-free rollback reflexes (breaker, safe-side refuge, streak-decay backoff). A rollback is always safe and a promotion is not, so only promotions wait for evidence. A new *virtual-probe* channel reconstructs counterfactual anchor errors to propose commits and to thin the probe tax ($12.5\% \to \approx 3\%$). Reliance on it is governed by an anytime-valid *validation e-process* (Prop. 1) that checks the model's predictions against real probes before trusting them. Against the watchdog, CUSUM and EWMA it trails on the benign suite by a small paired deficit. It ends the fewest episodes in net harm under the polarity fault and gives the largest duty-cycling reduction (Sec. 6). *(3) The cost of certification under matched controls.* The certified tier (restarted *mixture* e-processes (Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023), commit trials certifying the promoted object itself (Thm. 1), demotion, and the logged ledger of Prop. 3) costs attenuation under paired measurement (Sec. 6) and lowers no harm rate, and on the duty adversary it is statistically indistinguishable from the reflex tier. What it adds is a per-trial commit certificate, a tunable δ-budget, demotion, and a per-plant audit trail. This is, to our knowledge, the first matched-control accounting of a formal harm certificate's cost in this setting. *(4) Theory and negative results.* A blind-level and calibration analysis (Proposition A) covers Shewhart, CUSUM, EWMA and our own betting grid. Every fixed sensitivity tolerates persistent harm below its parameter indefinitely, and the charts need a null calibration the device lacks. We give a pathwise, variance-adaptive harm ledger that the gate logs and audits online (Prop. 3), a per-trial commit certificate for the promoted object itself (Thm. 1, which holds for adversarial adaptation under arbitrary model error, Prop. 2), the validation certificate (Prop. 1), and adoption (Thm. 2). A fully a-priori cumulative-harm bound remains an open problem (App. A.12). Negative results are reported too: a pre-registered blanket replacement of real probes by virtual ones failed, the uncertified adapter keeps the highest benign attenuation, and every measured harm-rate improvement, the duty-cycling case included, comes from the reflex tier. Certification therefore buys audits and guarantees.

2 Related Work

Adaptive filtering and ANC. FxLMS and variants remain the deployed standard (Morgan, 1980; Kuo & Morgan, 1996; Kajikawa et al., 2012). Instability under secondary-path error is the classical failure mode, mitigated by conservative step sizes, leakage, VSS heuristics (Kwong & Johnston, 1992; Akhtar et al., 2006), and online $\hat{S}$ re-estimation (Eriksson & Allie, 1989). Deep controllers (Zhang & Wang, 2021), learned update rules (Casebeer et al., 2021; 2023; Haubner et al., 2024) and offline RL filter selection (Shi et al., 2022; Luo et al., 2024) improve the average case without an off-distribution guarantee. The anchor/candidate structure itself is old. Two-path echo cancellation (Ochiai et al., 1977; Hänsler & Schmidt, 2004) adapts a background filter beside the live foreground filter and copies it across when a heuristic level comparison favors it. EGA replaces those heuristics with anytime-valid transfer/rollback tests at an explicit error budget and adds randomized probing, demotion and trust-gated model assistance. ANC virtual sensing reconstructs the error at microphone-less locations through identified paths (Moreau et al., 2008). Our virtual probes (Eq. 5) are the supervision-side counterpart. They evaluate a counterfactual filter at the real microphone, and the supervisor uses the result only as evidence.

Safe RL, shields, and monitored adaptation. Model-based certificates (Berkenkamp et al., 2017; Chow et al., 2018; Cheng et al., 2019), constrained policy optimization (Achiam et al., 2017) and safe exploration (surveys: García & Fernández, 2015; Brunke et al., 2022) assume plant knowledge.

Residual policy learning (Silver et al., 2018; Johannink et al., 2019) bounds the action but says nothing about the closed loop. Safe policy improvement (Ghavamzadeh et al., 2016; Laroche et al., 2019), Seldonian methods (Thomas et al., 2019) and conservative bandits (Wu et al., 2016; Kazerouni et al., 2017; Even-Dar et al., 2008) certify against a baseline offline or in expectation, whereas EGA certifies during deployment against a hot-swappable baseline by direct paired measurement. Runtime assurance with a trusted fallback is the Simplex idea (Seto et al., 1998; Sha, 2001). There, as in shielding (Alshiekh et al., 2018), safety filters (Wabersich & Zeilinger, 2021) and Recovery RL (Thananjeyan et al., 2021), the safety case rests on a model or critic verified at design time, the object that goes stale in our setting. EGA instead follows unfalsified and switching adaptive control (Safonov & Tsao, 1997; Stefanovic & Safonov, 2008; Hespanha et al., 2003) and builds its safety case from monitored measurements, with no plant model. Safeguarded learned optimization (Heaton et al., 2023) assumes checkable per-step progress, whereas we only have noisy paired measurements.

Anytime-valid and model-assisted inference. Our tests are betting-style e-processes (Ville, 1939; Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023; Howard et al., 2021), mixed over betting fractions and restarted at the floor as in e-detectors (Shin et al., 2024). Such tests already gate software rollouts (Johari et al., 2017; Lindon et al., 2022) and track deployed-model risk (Podkopaev & Ramdas, 2022). Virtual probes are an instance of model-assisted statistics, in which predictions add power while gold-standard measurements protect validity (Angelopoulos et al., 2023; Zrnic & Candès, 2024; Dudík et al., 2011; Jiang & Li, 2016). We apply the idea sequentially inside a control loop, where the model supplies predictions and real probes validate them.

Quickest change detection (QCD). Deciding when a monitored stream has turned bad is a mature subject (Page, 1954; Lorden, 1971; Moustakides, 1986; Basseville & Nikiforov, 1993; Veeravalli & Banerjee, 2014), and our diagnosis is known there too, as adaptive and mixture charts for unknown shift size (Sparks, 2000; Lai, 1998). Mixing over the sensitivity is therefore not new, and Proposition A claims no asymptotic separation from a horizon-tuned chart. What we add is (i) a blind-level accounting under the cumulative-harm functional a supervisor actually pays, covering Shewhart, CUSUM, EWMA and our own finite betting grid, (ii) the e-process route's analytic, distribution-free, anytime-valid threshold, as opposed to charts calibrated under a null that drifts with user, fit and session, and (iii) the closed-loop protocol (randomized paired probes, model-proposed but coin-certified commits, demotion to an invariant fallback), which a detector by itself does not give.

3 Safe Deployment-Time Adaptation

Plant. We instantiate the problem on single-channel feedforward ANC (block diagram in App. Fig. 2): a reference microphone measures noise x_t, the disturbance at the error microphone is $d = P * x$ through the primary path P, and a control FIR filter $W \in \mathbb{R}^L$ drives the loudspeaker, reaching the error microphone through the secondary path S (length L_S):

$$e_t \;=\; (P * x)_t + \big(S * \sigma(W * x)\big)_t + n_t, \tag{1}$$

with sensor noise n_t and driver saturation $\sigma(y) = \mathrm{clip}(y, \pm y_{\max})$. The inner adaptation is block FxNLMS with step size μ and leakage λ:

$$W \leftarrow (1-\lambda)W - \mu\,\frac{\sum_{t\in\text{blk}} e_t\,\hat{\boldsymbol{x}}_t^f}{\|\hat{\boldsymbol{x}}_{\text{blk}}^f\|^2 + \epsilon}, \qquad \hat{\boldsymbol{x}}^f = \hat{S} * x, \tag{2}$$

where $\hat{S}$ is a *stale* calibration estimate of S and $\hat{\boldsymbol{x}}_{\text{blk}}^f$ stacks the filtered reference over one H_b-sample block. Convergence requires the phase condition $\mathrm{Re}[S(\omega)\hat{S}^*(\omega)] > 0$ (Morgan, 1980). A re-seat perturbs (P, S) while $\hat{S}$ is fixed, so any $\mu > 0$ can diverge. Time is partitioned into control intervals of $n_I = MH_b = 256$ samples (16 ms, $M = 8$ blocks of $H_b = 32$ samples). At interval t an *adaptation policy* π chooses (μ_t, λ_t) from observable statistics.

The supervisor's problem. Let $\ell_t(\theta)$ be the error energy filter θ incurs at interval t. The exogenous (x, P, S) do not depend on the filter, so $\ell_t(\cdot)$ is well defined for all θ at once, though only the applied filter's loss is observed. The filter $\theta = 0$ is the *invariant fallback* (the feature off). A supervisor observes error energies, may apply any filter at any interval (hot-swap), and must guarantee, for an arbitrary, possibly adversarial π, that realized cumulative loss never substantially exceeds the fallback's, while retaining π's gains when it is good. Two facts make this tractable. Saturation and

Algorithm 1 EGA (one control interval)

base coin $\sim$ Bern$(1/K_t)$, $K_t \in \{8, 32\}$ (trust; 8 in alert), never paused; in trial an *extra* coin $\sim$ Bern$(1/P_{\mathrm{tr}}{=}1/5)$ (gain evidence only)—rates $\mathcal{F}_{t-1}$-measurable
if probe interval **then**
  snapshot lagged stale wealth $\log E_-^{\mathrm{stale}}$; target 0 (zero coin/cooloff/suspicion) $\succ W_{\mathrm{tr}}$ (trial, alternating) $\succ a$; certified bets (base coins, *fresh* pairing): harm on $\bar{Z}$, stale on $\bar{Z}^0 - \gamma_s$ if *armed* (bet consumes the flag); *mixture* wealths (Eq. 4); $V \mathrel{+}= \bar{Z}^2$; floor restart
  trial bets (*both* directions, one bet per realization: pair with stored opposite side, else store): gain on $-\bar{Z}^{\mathrm{tr}} - \gamma$; decrement trial budget; validate at anchor probes: two-sided bet on (prediction $-$ measurement) $\mp\gamma_v$; crossing $\Rightarrow$ distrust ($K_t \leftarrow 8$); 6 validated $\Rightarrow$ $K_t \leftarrow 32$
  if $E^{\mathrm{harm}} > 1/\delta_k$ **then**
    rollback ($W \leftarrow a$; refuge $W \leftarrow 0$ if $\log E_-^{\mathrm{stale}} > 1$); cooloff $C2^r$; $r{+}{+}$; alert (12 fast probes, trust intact); kill trial; disarm stale (anchor probe re-arms)
  else if $E^{\mathrm{stale}} > 1/\delta_k$ **then**
    demote: $a, W \leftarrow 0$; cooloff; $r{+}{+}$; distrust
  else if trial wealth $> 1/\delta_k$ **then**
    commit: $a \leftarrow W_{\mathrm{tr}}$; $r \leftarrow \max(0, r{-}1)$; reset gain/stale (harm wealth *continues*)
  end if
else
  run W; adapt by Eq. 2 unless cooloff; $W_{\mathrm{snap}} \leftarrow W$; virtual probe $\hat{e}(a)$ (Eq. 5): gain crossing (no cooldown) $\Rightarrow$ open trial, freeze $W_{\mathrm{tr}} \leftarrow W_{\mathrm{snap}}$; virtual harm >1 nat twice $\Rightarrow$ reflex rollback; breaker: rollback if error exceeds anchor level c_{brk} nats for p intervals
end if
trial dies at floor or $n_{\mathrm{tr}}{=}24$ bets (P_{cool} cooldown); $r \leftarrow \max(0, r{-}1)$ after 32 rollback-free scheduled probes (*streak decay*)

the sensor floor bound amplification, $\ell_t(\theta) \le B_{\max}\ell_t(0)$, and filters can be compared by *paired* measurements on adjacent intervals.

4 Evidence-Gated Adaptation

Randomized probes and paired evidence. EGA maintains a candidate W_t (updated by π through Eq. 2), an anchor a_t (initially 0), three test families and a trust-gated virtual-probe channel (Alg. 1). At every interval the gate flips its own *base* coin. With probability $1/K_t$, $K_t \in \{K, K_{\mathrm{slow}}\} = \{8, 32\}$ ($\mathcal{F}_{t-1}$-measurable, set by the trust machinery below, and fast also during a post-rollback *alert* window), the interval is a *probe*: the plant runs the anchor while the candidate is frozen. The base stream never pauses. During a *commit trial*, a second, independent coin at rate $1/P_{\mathrm{tr}} = 1/5$ adds extra trial probes (gain evidence only), so a proposal cannot open a certified-blind window. Data-independent coins keep probe times unpredictable, which blocks duty-cycling. Discarding the secondary-path ring-out ($\lceil L_S/H_b \rceil$ blocks), each probe yields $m = 4$ paired winsorized statistics against the preceding candidate interval,

$$Z_{j,b} \;=\; \mathrm{clip}\big(\log \ell^{(b)}(\mathrm{cand}) - \log \ell^{(b)}(\mathrm{anchor}),\, \pm c\big), \tag{3}$$

whose block-mean $\bar{Z}_j = m^{-1}\sum_b Z_{j,b}$ is the probe's single betting statistic for the harm test (a false rollback is conservative). Gain evidence comes from commit trials (below). Certified bets ride the base coins (running through trials) and fire only when the preceding interval ran the candidate (*fresh pairing*, with a gap of one interval, so consecutive probes cannot re-bet stale blocks). The pairing is pre-probe only. The measured candidate interval predates the probe coin, so no process can shape what is sampled. No measured windowed drift in any suite or state reaches $\gamma/2$ (App. A.10). Each probe is independently a *zero probe* w.p. $1/J$ (a second gate coin, and probes during cooloff and on suspicion also run $\theta = 0$), producing $\bar{Z}_j^0$ comparing the anchor to passive ($J = 4$ deployed).

Mixture betting e-processes with restarts. Each family $f \in \{\mathrm{harm}, \mathrm{gain}, \mathrm{stale}\}$ tracks a nonnegative wealth, betting *once per probe* on the block-mean, using a uniform *mixture* over a grid $\mathcal{H}_f$ in place of a single fraction. For harm (H_0: $\mathbb{E}[\bar{Z}_j \mid \mathcal{F}_{j-1}] \le 0$, candidate no worse than anchor):

$$E_t \;=\; \frac{1}{Q}\sum_{\eta\in\mathcal{H}_f}\prod_{j\le t}\big(1 + \eta\,\bar{Z}_j\big), \qquad \mathcal{H}_{\mathrm{harm}} = \mathcal{H}_{\mathrm{stale}} = \{0.005, \ldots, 0.2\}, \tag{4}$$

($Q = 6$ log-spaced points, 5 for gain, up to 0.35), restarted at 1 whenever the mixture would fall below it. A convex combination of nonnegative supermartingales is again a nonnegative supermartingale, so validity is unaffected. A single fraction η, by contrast, tolerates harm below its blind level $b_-\eta v_{\mathrm{lo}}$ ($b_-{=}0.27$, $v_{\mathrm{lo}}{=}1.596$) indefinitely. The grid shrinks that level roughly $40\times$, to ≈ 0.002 nats, still nonzero (Proposition A). Restarts are stopping times of the mixture. With $\delta_k \propto \delta/(k{+}1)^2$ spent over restarts, Ville's inequality gives family-wise validity (jointly $\le \delta$, Lem. 2). The gain test bets on $-\bar{Z}_j^{\mathrm{tr}} - \gamma$ (trial object better by margin γ, below). Staleness bets on $\bar{Z}_j^0 - \gamma_s$ ($\gamma_s = 0.25$, a margin sized so typical measured reference-gap drift cannot fire demotion alone, measured levels

in App. A.10). Each staleness bet consumes its stored anchor reference, one bet per realization, re-armed only by a fresh anchor probe.

Events. *Rollback* (harm test fires) sets $W \leftarrow a$ and freezes adaptation for $C \cdot 2^r$ probe cycles, $r \mathrel{+}= 1$ (uncapped). Rollbacks carry no certificate, so they pick the safe side. Lagged ($\mathcal{F}$-measurable) staleness wealth over 1 nat retargets $W \leftarrow 0$ (*safe-side refuge*), and the rollback opens a 12-probe fast-rate *alert* without voiding trust. *Commit* (a trial's wealth crosses) sets $a \leftarrow W_{\mathrm{tr}}$, the frozen trial object that every trial bet measured, so the certified object and the promoted object coincide (the ratchet), and $r \leftarrow \max(0, r-1)$. Backoff decays one step per commit and one per 32 rollback-free scheduled probes (*streak decay*) and never resets. Within a burst, cooloffs double and the burst ends after $\log_2$-many rollbacks, while isolated benign rollbacks stop escalating. Gain and staleness restart at commits. The harm wealth continues across commits, since its null is stated per bet against whichever anchor is in force, so a forced commit cannot refresh an adversary's threshold budget. *Demotion* (staleness test fires) sets $a, W \leftarrow 0$. This re-grounds the chain in the invariant fallback when the plant itself changes under a good anchor. Demotions and validation crossings (not ordinary rollbacks) revoke trust.

Model-assisted probing: virtual probes, commit trials, trust. Probes buy paired evidence at a standing attenuation tax of $1/K$. With a linear plant and an exogenous environment, the device can reconstruct what the anchor would have measured during a candidate interval:

$$\hat{e}(a) \;=\; e \;-\; \hat{S} * \sigma(W * x) \;+\; \hat{S} * \sigma(a * x), \tag{5}$$

whose error scales with the filter difference $(S-\hat{S})*(\sigma(W*x)-\sigma(a*x))$ and not with the signal level. The statistic is nearly exact while the calibration holds (bias median 0.007, p95 0.072 nats before a re-seat) and useless once the model is wrong (p95 0.81 after a re-seat, median -1.33 under polarity). A pre-registered blanket replacement of real probes failed (pooled p95 0.64), so virtual probes serve only the trust-gated roles those strata support. Virtual evidence never enters a certificate. The model can only propose, and certified decisions are taken on real, coin-scheduled probes. *(i) Proposal → trial:* A virtual gain wealth accumulates every candidate interval. A crossing freezes a copy W_{tr} of the proposed snapshot and opens a *commit trial*. Probes at rate $1/P_{\mathrm{tr}} = 1/5$ alternate W_{tr} with the anchor, and the gain family bets in both directions, one bet per realization (a measurement pairs with the stored opposite side or is stored, never both), so every bet compares the same two gate-stored objects and a crossing certifies the object that is promoted (Thm. 1). The live candidate keeps adapting at no freeze cost. A trial dies at its floor or after $n_{\mathrm{tr}} = 24$ bets (192-interval cooldown). *(ii) Reflex:* Virtual harm >1 nat on consecutive intervals triggers a certificate-free rollback, since a false positive is safe. *(iii) Trust:* A two-sided validation e-process bets on the gap between the virtual prediction and each real anchor probe (margin $\gamma_v = 0.25$). Six validated probes thin the rate to $K_{\mathrm{slow}} = 32$ (tax $12.5\% \to \approx 3\%$). Validation crossings and demotions restore $K_t = 8$, while ordinary rollbacks trigger only the alert. Validation has the most power where the model is worst (polarity gap $+1.64 \pm 1.06$ nats, so trust goes within a few probes, and post-fault $+0.63 \pm 1.17$). The weak case, benign re-seats ($+0.03 \pm 0.77$), is covered by the K_{slow} floor and the trial requirement (Props. 1–2).

Circuit breaker. Divergence can outrun probe-resolution tests, so EGA also rolls back when the candidate's noise-normalized error exceeds the anchor's recent level by c_{brk} nats for p consecutive intervals. This is a threshold rule. We use it only for retreat, where a blind spot costs nothing (Proposition A).

The adapters. EGA accepts any proposer, and we train two adapters observing 25 features (signal energies, adaptation and gate state), acting every 16 ms with $(\log_{10}\mu, \lambda)$, meta-trained across randomized acoustics and noise (Tobin et al., 2017) through the gate, whose events are nondifferentiable: *EGA-RL*, a PPO MLP (Schulman et al., 2017), and *EGA-GRU*, a GRU trained by BPTT through the differentiable simulator (the Meta-AF recipe (Casebeer et al., 2023)). Training through the gate matters, since wrapping a separately trained adapter after the fact costs attenuation (4.6 vs. 3.6 dB).

5 Theory

Full statements and proofs are in App. A. Probes are Bernoulli coins at a predictable three-rate schedule (trust tiers $1/K_t$ plus the trial rate, certified bets ride non-trial coins), each yielding the block-mean statistic $\bar{Z}_j \in [-c, c]$ over m paired winsorized blocks. Every certified statement

concerns this statistic. Writing ς_t for the applied system's block-mean winsorized log-excess over the anchor chain, each sampled $\bar{Z}_j$ is a paired estimate of ς_t at the pre-probe live interval (App. A: block-balance for the interval-level transfer, not claimed unconditionally). A monitor sees $\{\bar{Z}_j\}$ and picks intervention times. Deployed monitors (watchdog, CUSUM, EWMA, a single-fraction e-process, and our mixture) share a sensitivity level ϑ with non-positive firing-statistic drift below it. We ask how low validity lets that level sit, and what calibration costs.

Validity (informal, Lem. 2). Every betting family, the uniform mixture over its grid and any $\mathcal{F}_{t-1}$-measurable choice of predictable coins, is a nonnegative supermartingale under its null, so each errs, anywhere in time, w.p. at most its spending total (Ville). For $\delta \in (0, 1)$, the three e-test families each spend $\delta/3$ of a shared restart-indexed schedule $\delta_k \propto \delta/(3(k{+}1)^2)$, jointly $\leq \delta$. The validation family spends an independent $\delta_v = 2\delta/3$ (Prop. 1). It is accounted separately, since it certifies the evidence channel and not a filter. Predictable selection affects which evidence is gathered (Lem. 4) but not validity.

Proposition A (Blind levels and calibration, informal, App. A.5)**.** *Under a constant adversary holding h below a monitor's blind level ϑ, the firing statistic has non-positive drift, and once the threshold is validity-calibrated,* $\mathrm{Harm}(T) \geq \frac{3}{4}hT$. *When validity forces $\vartheta \geq c$ (a memoryless threshold on a clipped statistic with a clip atom, the deployed watchdog, measured), it never fires and* $\mathrm{Harm} = cT$. *Our grid's blind level $b_-\eta_{\min}v_{\mathrm{lo}} \approx 0.002$ nats (v_{lo}=1.596, distinct from the upper bound $\bar{v}$=1.81 used below) sits $40\times$ below the largest single fraction but is still nonzero, and the measured worst case sits at it. Above it, worst-case harm is capped at $\tilde{O}(\max\{\bar{v}\eta_{\min}T, \sqrt{\bar{v}T}\})$, a rate a horizon-tuned CUSUM also attains. On a device the two routes differ in calibration. The e-process threshold $\log(1/\delta_k)$ is analytic, distribution-free and anytime-valid, whereas the charts must be simulation-calibrated under a null the device does not have.*

The blind spot is visible in the loop too. The watchdog that edges our reflex tier on benign medians defends nothing below its level over deployment hours (duty adversary, Sec. 6), and above that level it fires at $2.0\times$ our rollback rate.

Assumption 1 (Bounded amplification, holds w.h.p. by construction)**.** *Driver saturation and the sensor-noise floor give $\ell_t(\theta) \leq B_{\max}\,\ell_t(0)$ for all reachable θ and all $t \leq T$, except w.p. $Te^{-n_I/16}$ (Lem. 1). $B_{\max}$ is a coarse backstop from the benchmark's physical constants. The operational cap is the breaker's $e^{c_{\mathrm{brk}}}$, which every configuration (both tiers) runs.*

Proposition 1 (Validation certificate, certified trust)**.** *The validation family is an anytime-valid test of the virtual-probe channel's calibration: under the null that virtual predictions match real anchor measurements within γ_v in winsorized conditional mean, it errs anywhere in time w.p. $\leq \delta_v = 2\delta/3$ at the analytic threshold $\log(1/\delta_k)$: every distrust validly rejects the channel. All model influence routes through $\mathcal{F}_{t-1}$-measurable decisions (probe rate and trial proposals, trust-gated, and the virtual-harm reflex, deliberately ungated since a false rollback is safe). This is the certificate that governs the reflexes-plus-virtual-probes configuration, the headline system evaluated in Sec. 6 (App. A.7). The optional filter-level tier below adds its own, separate certificates.*

5.1 The certified tier and its cost

Enabling the harm/gain/staleness e-tests (Alg. 1) adds three guarantees at the shared δ above, stated informally here. Full statements and proofs are in App. A under "the certified tier, optional apparatus".

Assumption 2 (Paired conditional drift)**.** *For adjacent intervals and fixed θ, $|\mathbb{E}[\log \ell_{t+1}(\theta) - \log \ell_t(\theta) \mid \mathcal{F}_{t-1}]| \leq \kappa < \gamma/2 = 0.25$. Only the results of this subsection consume κ.*

Remark (measured). The bound concerns the conditional mean. Realized fluctuations are charged by the ledger's V_n below. Measured per suite $\times$ filter state, fault suites included: medians ≤ 0.019, worst-cell p95 0.070, max 0.176, and no window reaches $\gamma/2$ anywhere (App. A.10). A fraction $\nu \leq 0.7\%$ exceeds a stricter consumed level (Assumption 2′), budgeted wherever quoted.

Theorem 1 (Per-trial commit certificate, informal)**.** *Fix any adaptation process π (adversarial, aware of EGA's internals but not the gate's future coins) and any model $\hat{S}$. A commit trial freezes a copy W_{tr} of the proposed snapshot and bets, once per realization, on real paired probes of the fixed pair (W_{tr}, a). Its wealth is a fresh restarted e-process at its own δ_k. On the validity and trial-concentration events (App. A.11, jointly $\geq 1 - \delta/3 - \delta_g$ over all trials, δ_g an auxiliary concentration parameter set*

by the analyst, not drawn from δ), every commit certifies that the promoted object improved on the outgoing anchor by at least $\gamma - \kappa_{\mathrm{tr}}$ nats in run-average winsorized conditional mean, where κ_{tr} is the in-trial pairing-gap drift (Assumption 2). The certificate is pairwise and per event. It attaches to the object installed and is not chained across commits (App. A.12).

Remark (measured margin). At the proof's high-probability gap constant, the margin is positive under typical (median) drift ($\gamma - \kappa_{\mathrm{tr}} \approx 0.24$ nats, $\approx$1.1 dB) but not at p95, where $\kappa_{\mathrm{tr}} > \gamma$ and a commit certifies no margin, and protection then comes from the harm ledger and validation certificate together with the reflexes (App. A.10).

Proposition 2 (Certificate independence of the model)**.** *Fix any $\hat{S}$ (arbitrarily wrong) and any π. The virtual-probe channel alters only which intervals are probed (rate, proposals and alert window are $\mathcal{F}_{t-1}$-measurable, harm and staleness bet only on non-trial coins, and trials certify gate-stored frozen objects). The validity of every betting family, the per-trial commit certificate above and the logged ledger below hold verbatim under arbitrary model error. The model changes which intervals get sampled and how many of them. That performance cost is logged, and validity is untouched (App. A.6).*

Proposition 3 (Logged harm ledger, pathwise, audited online)**.** *Within any betting run of the harm family (between restarts), let n be the sampled-bet count, $V_n := \sum_j \bar{Z}_j^2$ the realized quadratic variation, and $A'_k := \log(1/\delta_k) + \log Q$ the run's threshold. On every prefix on which the mixture has not fired, pathwise and deterministically,*

$$\textstyle\sum_{j\le n} \bar{Z}_j \;\le\; \min\big\{\, c\,n,\;\; \eta_{\mathrm{eff}}\, V_n + \frac{Q}{S_{\mathcal{H}}} A'_k,\;\; \min_{\eta\in\mathcal{H}} \big[\frac{A'_k}{\eta} + \eta\, V_n\big]\big\},$$

with grid mass $S_{\mathcal{H}} := \sum_{\eta\in\mathcal{H}} \eta$ and η_{eff} a grid-determined constant (App. A). Thresholds are paid only at event-terminated runs (never at commits), and floor restarts cost $\log Q$ only at single-η readings. Outside the staleness-suspicion zero-probe redirect (certified-tier only, gated on the e-tests, so the reflex tier never opens it and the sampling hypothesis holds there unconditionally, App. A), the sampled statistics represent the applied system's winsorized excess at the predictable per-tier probe rates. The gate logs both sides ($\sum_j \bar{Z}_j$, V_n, N_{samp}, restart/event counts, and the suspicion-redirect occupancy), and the ledger is auditable online, per plant (Fig. 1).

Remark (deployed reading). The allowance rate is $\eta_{\mathrm{eff}}\bar{v} = 0.25$ nats/probe against the trivial $c = 2$ (2.2$\times$ below the range c^2, App. B.4). The suspicion-redirect occupancy excluded is 0.0% mean / 0.0% p95 (certified tier, logged) vs. 0% for the reflex tier, which never opens the redirect. Rare-restart, rare-rollback runs read $2\sqrt{(N_{\mathrm{ev}}+1)A'_k V_T} = \tilde{O}(\sqrt{T})$.

What is not claimed. A fully a-priori cumulative-harm account (bounding event counts and transferring the sampled sum to all intervals a priori) requires additional assumptions (piecewise stationarity, drift transfer across reference gaps) under which the assembled bound exceeds the trivial cap cT at every deployed horizon. We therefore certify the logged ledger and the per-event certificates above, and report the a-priori route as an open problem (App. A.12). The drift measurements characterize the pairing margins empirically (App. A.10). Raw energy stays capped throughout by Assumption 1 and, within p intervals, the breaker.

Theorem 2 (Adoption and ratchet, informal)**.** *Since commits are exclusively trial-certified, adoption is two-phase. A calibrated virtual channel (Prop. 1) proposes a trial once a candidate's improvement exceeds margin $\beta > \gamma$. The trial then crosses within N real paired bets on the fixed pair w.p. $\geq 1 - \delta - \delta'$, with N and its per-bet drift ε_0 given in App. A.13. The committed gain persists until the plant changes again.*

Remark (deployed reading). At deployed constants, a decisively better candidate crosses in $\approx$15 bets ($\approx$4.0 s, decoupled from K_t). For marginal candidates the bound fails its own hypothesis ($N \approx 100 > n_{\mathrm{tr}} = 24$, App. A.13).

6 Experiments

Benchmark. We built a GPU-vectorized feedforward-ANC environment (16 kHz, 1024 parallel plants, $\sim$58k intervals/s on one GPU) with randomized causal primary paths, minimum-phase secondary paths with latency, and $\hat{S}$ at 30 dB SNR (median causal-Wiener ceiling $\sim$17 dB). Noise

Table 1: Main suite (re-seats, 512 held-out plants, 3 seeds pooled): attenuation median/p10/final-2 s (dB, 95% CI), divergence-flag and net-harm rates (gated div% flags are probe hand-off transients). Bottom: *matched* arms, same in-loop checkpoints, only the supervisor varies. †our configurations.

Method	med	p10	final	div%	harm%
Passive ($W \equiv 0$)	-0.0 $\pm$0.0	-0.0	-0.0	0.0	0.0
Near-frozen (μ=10^{-4})	1.6 $\pm$0.2	0.5	1.9	5.9	1.6
FxNLMS μ=0.005	4.7	0.3	4.0	31.6	8.2
FxNLMS μ=0.05	2.6	-3.8	1.0	76.4	24.8
FxNLMS μ=0.2	-2.7	-11.0	-5.5	97.7	70.5
Oracle fixed-μ (per plant)	5.1	0.4	–	–	–
VSS-FxNLMS	2.7 $\pm$0.2	0.8	2.7	7.8	1.0
FxNLMS μ=0.02 + online $\hat{S}$	1.0 $\pm$0.3	-1.7	1.5	46.1	27.0
GRU adapter (Meta-AF–style)	6.4 $\pm$0.4	2.4	5.4	3.5	0.8
PPO (no gate)	6.3 $\pm$0.4	1.8	5.2	11.5	2.1
Watchdog + GRU (post hoc)	5.9 $\pm$0.4	1.4	4.8	4.1	2.1
EGA + GRU (wrapped post hoc)	3.6 $\pm$0.2	-0.1	2.7	20.3	8.0
Watchdog (matched ckpt)	5.9 $\pm$0.3	1.7	4.9	4.5	0.8
CUSUM (matched, null-calib.)	5.9 $\pm$0.4	1.6	4.9	4.3	0.2
EWMA (matched, null-calib.)	5.9 $\pm$0.4	1.6	4.9	4.7	0.4
EGA no-VP (matched)	4.1 $\pm$0.2	1.4	3.6	7.4	1.2
EGA-RL (ours)†	4.6 $\pm$0.3	1.0	3.7	11.1	2.7
EGA-Full (ours)†	4.6 $\pm$0.2	1.2	3.8	7.8	2.0
EGA-Reflex+VP (ours)†	5.7 $\pm$0.4	1.3	5.0	5.3	1.0

regimes switch (engine, street, HVAC, and babble, which is held out) with ± 6 dB drift. Re-seats perturb (P, S) under a 300 ms crossfade, and $\hat{S}$ stays stale. Episodes run 12 s test / 6 s train, 1–2 re-seats (simulator verified against a scalar reference, App. B).

Protocol and matched arms. 512 held-out plants per condition (training-disjoint seeds) across 8 suites: stationary and re-seat (in-distribution), OOD noise (babble), OOD severity (2–4 re-seats), $\hat{S}$ at 10/0 dB SNR, *polarity-inverted* $\hat{S}$ (a wiring/firmware sign fault), and *fault_mid*, the same fault striking at the first re-seat. Metrics: episode attenuation $10\log_{10}(\sum d^2 / \sum e^2)$ (median, p10), final-2 s attenuation, *divergence* (error $>2\times$ disturbance on $>2\%$ of intervals), *net harm* ($\sum e^2 > 1.05 \sum d^2$: worse than ANC off). Baselines: true passive ($W \equiv 0$), near-frozen ($\mu = 10^{-4}$), an FxNLMS grid $\mu \in [0.005, 0.2]$ and its per-plant oracle fixed μ, VSS-FxNLMS (Kwong & Johnston, 1992), hysteresis (Hespanha et al., 2003), Eriksson online-$\hat{S}$ (Eriksson & Allie, 1989), and uncertified Meta-AF–style GRU (Casebeer et al., 2023) and PPO. We never compare mechanisms across checkpoints. The *matched* arms wrap the same in-loop checkpoints (3 seeds, pooled) on the same plants. Tier-0 is **Watchdog** (breaker-only veto, anchor pinned at passive) plus **CUSUM/EWMA** on the same residual, null-calibrated offline (stationary suite, watchdog's false-alarm budget). Tier-1 is **EGA-Reflex+VP** (reflexes + trust-gated virtual probes, e-tests off). Tier-2 is **EGA-Full**, and we also run **EGA no-VP** (virtual evidence off, commit path kept via scheduled proposals). Rows wrapping the separately trained GRU are labeled *post hoc*. All matched arms share checkpoints trained under the full (tier-2) gate, a train/deploy mismatch for the reduced arms that the on-policy comparison below resolves. Medians carry 95% cluster-bootstrap CIs by plant. Matched differences are paired.

Cost of each supervision tier. Tier 1 (reflexes + trust-gated virtual probes, no e-tests) reaches 5.7 ± 0.4 dB median vs. the matched watchdog's 5.9 ± 0.3 (paired -0.13 $[-0.16, -0.11]$, CI excludes zero). CUSUM 5.9 and EWMA 5.9 are comparable. The certified tier (trials, commits, demotion, Prop. 3's ledger) costs $+0.92$ dB paired (excludes zero, medians differ by 1.16 dB). The no-VP arm (4.1 dB) keeps the commit path via scheduled proposals, which isolates virtual evidence (paired effect $+0.62$ dB, excludes zero). The uncertified GRU keeps the highest raw decibels (6.4 ± 0.4). To remove the mismatch, three native recipes (reflex-tier gate in-loop, full gate in-loop, ungated watchdog-native), three seeds each, are evaluated on two batches of 128 held-out plants with paired per-plant differences and cluster-by-plant bootstrap CIs (`onpolicy_pricing.py`, supplementary). Tier 1 reaches 5.6 dB vs. the watchdog-native's 5.8 dB (paired -0.14 dB $[-0.19, -0.11]$), but the order reverses under polarity, -1.6 vs. -2.4 dB (paired +0.70 dB $[+0.60, +0.81]$, a win). Certification costs +0.94 dB paired $[+0.82, +1.03]$ more (+0.41 dB fault_mid, -0.01 dB polarity).

Calibration-fault suites. Polarity-inverted calibration turns every blind adapter into an amplifier (FxNLMS -7.5 dB median, PPO -6.4, GRU -3.9). All supervised arms cap it at comparable raw decibels (tier-1 -1.8 vs. EWMA -1.6, CUSUM -1.8, watchdog -2.4 dB), and only the net-harm flags separate the arms. Every EGA arm ends fewer episodes in net harm than every chart, at both

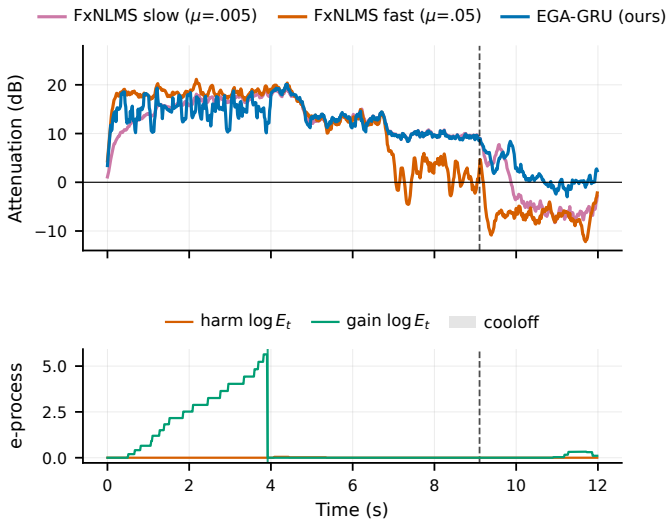


Figure 1: Held-out re-seat episode (re-seat dashed): fast FxNLMS diverges. EGA adapts, rolls back within a probe cycle, and re-commits. The certified tier's ledger-audit figure is in App. C (Fig. 3).

thresholds (harm_{05} and harm_{25} per arm, Tab. 2). Validation (gap $+1.64 \pm 1.06$ nats) withdraws trust within a few probes. The residual harm is capped but nonzero (App. C).

Duty-cycling adversary. A proposer that spikes inside the breaker's patience window leaves the reflex tier with 12.3% severe-harm episodes. The like-for-like comparators, budget-matched charts fed the same virtual evidence that the reflex tier's harm reflex uses, still reach 21.4/21.2% (about 1.7–1.7× higher at matched rollback rates). Plain charts and the watchdog do worse still (38.3/45.6/40.2%, ungated 70.9%, listed for context in App. C). The full tier is statistically indistinguishable from the reflex tier (12.2% vs. 12.3%, CIs $[10.5, 13.8]$/$[10.7, 14.0]$ overlap).

Duty-cycling under a calibration fault. With the duty adversary run under the polarity fault ($\hat{S}$ wrong from t=0, where virtual evidence is least trustworthy, App. A.10), the reflex tier still ends only 57.0% of episodes severely net-harmful, against 85.9/94.1/85.9% for watchdog/CUSUM/EWMA (full tier 56.2%). When the model fails, real-probe reflexes and backoff preserve the advantage. Feeding the untrusted stream barely helps CUSUM and only partly helps EWMA (App. C).

Fault at a committed anchor (fault_mid). Blind adapters amplify (FxNLMS ≈ -4.7 dB median, oracle -4.6, GRU -0.7). Tier 1 and the watchdog tie near passive (0.9 vs. 0.9 dB). The charts split around that pair (CUSUM 1.0 dB narrowly ahead, EWMA 0.8 dB narrowly behind). The full tier ends at 0.1 dB, above every blind adapter but below tier 1. Its pre-fault anchor persists until refuge, and demotion never fires (0.00/ep.).

Long horizon (600 s). At 600 s ($n = 24$ plants), tier 1 holds 4.54 dB vs. the watchdog's 5.19 (CUSUM 4.96, EWMA 5.16), with overlapping CIs and comparable rollback rates (6.71 vs. 5.54–8.38/ep.). The full tier falls to 2.38 dB (premium 2.16 dB). Grid-allowance compliance and harm-vs-δ flatness (Tab. 3) are in App. C.

7 Limitations and Conclusion

Several limitations remain. The 600 s comparison rests on $n = 24$ plants, the ablations use one seed, and the drift and virtual-probe-bias assumptions are measured only in simulation. The certificates concern winsorized energy (App. A.3), and the theory establishes calibration, without a separation between monitor classes. Where parameters can be hot-swapped and error is measurable, monitoring alone supervises adaptation without a plant model. The reflex tier produces the harm-rate and attenuation results, and the certified tier buys guarantees and an audit trail at a measured paired cost.

Ethics Statement

This work aims to make on-device adaptive audio systems safer for users while keeping their attenuation. The same supervisor design could apply to other consumer signal chains that run learned controllers, though we do not evaluate that here. All experiments are simulations, with no human subjects or personal data. We do not foresee negative societal impacts specific to this method beyond those generic to improved consumer audio technology.

Reproducibility Statement

The complete benchmark environment, gate, training and evaluation code, and all experiment configurations are included as anonymized supplementary material, with fixed, disjoint seed streams for training and evaluation (App. B). The vectorized simulator is cross-validated against a scalar textbook implementation and an exact causal-Wiener optimum. Every number and table in the paper is generated programmatically from the evaluation outputs. Full proofs of all claims are in App. A. Reproducing all results takes under 12 GPU-hours on one GPU.

AI Use Statement

The authors did not use large language models or other generative AI tools in the research or preparation of this manuscript.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Muhammad Tahir Akhtar, Masahide Abe, and Masayuki Kawamata. A new variable step size LMS algorithm-based method for improved online secondary path modeling in active noise control systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):720–726, 2006.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, 2018.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ, 1993.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Jonah Casebeer, Nicholas J. Bryan, and Paris Smaragdis. Auto-DSP: Learning to optimize acoustic echo cancellers. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

Jonah Casebeer, Nicholas J. Bryan, and Paris Smaragdis. Meta-AF: Meta-learning for adaptive filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:355–370, 2023.

Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI Conference on Artificial Intelligence*, 2019.

Yinlam Chow, Ofir Nachum, Edgar Duéñez-Guzmán, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Stephen J. Elliott and Philip A. Nelson. Active noise control. *IEEE Signal Processing Magazine*, 10 (4):12–35, 1993.

Larry J. Eriksson and Mark C. Allie. Use of random noise for on-line transducer modeling in an adaptive active attenuation system. *The Journal of the Acoustical Society of America*, 85(2): 797–802, 1989.

Eyal Even-Dar, Michael Kearns, Yishay Mansour, and Jennifer Wortman. Regret to the best vs. regret to the average. *Machine Learning*, 72(1-2):21–37, 2008.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Mohammad Ghavamzadeh, Marek Petrik, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Eberhard Hänsler and Gerhard Schmidt. *Acoustic Echo and Noise Control: A Practical Approach*. Wiley, Hoboken, NJ, 2004.

Thomas Haubner, Andreas Brendel, and Walter Kellermann. End-to-end deep learning-based adaptation control for linear acoustic echo cancellation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:227–238, 2024.

Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded learned convex optimization. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 7848–7855, 2023.

João P. Hespanha, Daniel Liberzon, and A. Stephen Morse. Hysteresis-based switching algorithms for supervisory control of uncertain systems. *Automatica*, 39(2):263–272, 2003.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

Yoshinobu Kajikawa, Woon-Seng Gan, and Sen M. Kuo. Recent advances on active noise control: open issues and innovative applications. *APSIPA Transactions on Signal and Information Processing*, 1:e3, 2012.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Sen M. Kuo and Dennis R. Morgan. *Active Noise Control Systems: Algorithms and DSP Implementations*. Wiley, 1996.

Raymond H. Kwong and Edward W. Johnston. A variable step size LMS algorithm. *IEEE Transactions on Signal Processing*, 40(7):1633–1642, 1992.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning (ICML)*, 2019.

Michael Lindon, Chris Sanden, and Vaché Shirikian. Rapid regression detection in software deployments through sequential testing. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Zhengding Luo, Haozhe Ma, Dongyuan Shi, and Woon-Seng Gan. GFANC-RL: Reinforcement learning-based generative fixed-filter active noise control. *Neural Networks*, 180:106687, 2024.

Danielle Moreau, Ben Cazzolato, Anthony Zander, and Cornelis Petersen. A review of virtual sensing algorithms for active noise control. *Algorithms*, 1(2):69–99, 2008.

Dennis R. Morgan. An analysis of multiple correlation cancellation loops with a filter in the auxiliary path. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):454–467, 1980.

George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.

Kazuo Ochiai, Takashi Araseki, and Takasi Ogihara. Echo canceler with two echo path models. *IEEE Transactions on Communications*, 25(6):589–595, 1977.

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

Tristan Pepin, Nicolas J. Pignier, and Jean Rabault. Deep reinforcement learning applied to active noise control: a smart, adaptive, variable step size FxLMS algorithm design. Preprint, 2021.

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

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

Michael G. Safonov and Tung-Ching Tsao. The unfalsified control concept and learning. *IEEE Transactions on Automatic Control*, 42(6):843–847, 1997.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Danbing Seto, Bruce Krogh, Lui Sha, and Alongkrit Chutinan. The Simplex architecture for safe online control system upgrades. In *American Control Conference (ACC)*, pp. 3504–3508, 1998.

Lui Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001.

Dongyuan Shi, Bhan Lam, Kenneth Ooi, Xiaoyi Shen, and Woon-Seng Gan. Selective fixed-filter active noise control based on convolutional neural network. *Signal Processing*, 190:108317, 2022.

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.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Ross S. Sparks. CUSUM charts for signalling varying location shifts. *Journal of Quality Technology*, 32(2):157–171, 2000.

Margareta Stefanovic and Michael G. Safonov. Safe adaptive switching control: Stability and convergence. *IEEE Transactions on Automatic Control*, 53(9):2012–2021, 2008.

Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3): 4915–4922, 2021.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Venugopal V. Veeravalli and Taposh Banerjee. Quickest change detection. In *Academic Press Library in Signal Processing*, volume 3, pp. 209–255. Elsevier, 2014.

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

Kim P. Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021.

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.

Bernard Widrow and Samuel D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, 1985.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning (ICML)*, 2016.

Hao Zhang and DeLiang Wang. Deep ANC: A deep learning approach to active noise control. *Neural Networks*, 141:1–10, 2021.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *International Conference on Machine Learning (ICML)*, pp. 62993–63010, 2024.

A PROOFS

A.1 NOTATION AND PRELIMINARIES

Control intervals are indexed by $t = 1, 2, \ldots$. Each contains M blocks of H_b samples. For a filter θ, $\ell_t(\theta)$ denotes the error energy that would be incurred running θ throughout interval t, and $\ell_t^{(b)}(\theta)$ the block-level energies. The exogenous environment process $\xi = (x, P, S, n)$ does not depend on the applied filter, so $\{\ell_t(\theta)\}_\theta$ are simultaneously well defined. $\mathcal{F}_t$ is the σ-algebra generated by the exogenous processes through interval t together with all adapter randomness used through t and the supervisor's coins through t. Hence the counterfactual losses $\ell_t(\theta)$ are $\mathcal{F}_t$-measurable for every fixed θ, and the gate's interval-$(t{+}1)$ coins are independent of $\mathcal{F}_t$.

Two randomized coin streams at predictable rates, commit trials. At each interval the gate draws a *certified base coin* $G_t \sim \mathrm{Bernoulli}(1/K_t)$ from its own generator. The coin's outcome is independent of $\mathcal{F}_{t-1}$ and of all environment and adapter randomness, while the rate $1/K_t$, $K_t \in \{K, K_{\mathrm{slow}}\} = \{8, 32\}$, is $\mathcal{F}_{t-1}$-measurable: the trust-set tier of Sec. 4, with the fast tier K also pinned during a post-rollback *alert window* (the $a_{\mathrm{al}} = 12$ scheduled probes after any rollback-like event). Trust itself is voided only by validation crossings and demotions. The base stream never pauses: during an active *commit trial* (below) the gate draws a second, independent coin at rate $1/P_{\mathrm{tr}} = 1/5$, and intervals where the trial coin fires but the base coin does not are *extra* trial probes. A proposal adds gain evidence but can never suspend certified sampling. *Certified coins* are the base draws ($G_t = 1$): only these carry harm and staleness bets. Trial draws feed the gain family only. On each probe the gate draws a further independent coin $\zeta_t \sim \mathrm{Bernoulli}(1/J)$: $\zeta_t = 1$ makes it a *zero probe*, running the invariant fallback $\theta = 0$. Zero-probe times are $\mathrm{Bernoulli}(1/(K_t J))$ among certified coins at every interval, and are as unpredictable to the adaptation process as the probe times themselves. (Probes during cooloff and under staleness suspicion also run $\theta = 0$, and those rules are predictable.) *Commit trials.* When the virtual gain wealth crosses its threshold (an $\mathcal{F}_{t-1}$-measurable trigger), the gate freezes a copy $W_{\mathrm{tr}} := W_{\mathrm{snap}}$ of the proposed snapshot and opens a trial. Trial-coin probes alternate between W_{tr} and the anchor (zero probes take precedence and do not advance the alternation). Base-coin probes run the anchor or the fallback as usual and carry their certified bets. The gain family bets in both directions under a strict *one-bet-per-realization* rule: a probe's measurement either pairs with a stored opposite-side measurement, in which case that single bet consumes both and neither enters any later bet, or is stored for the next opposite-side probe, never both ($\approx n/2$ bets per n non-zero trial probes). The live candidate keeps adapting throughout. A commit promotes W_{tr}. The trial dies at its wealth floor or after $n_{\mathrm{tr}} = 24$ bets, with a proposal cooldown of $P_{\mathrm{cool}} = 192$ intervals before the next trial. Trial draws are data-triggered evidence for the *gain family only*, never the harm or staleness certificate. *Fresh pairing and arming.* The harm and staleness families bet only when the immediately preceding interval ran the candidate, so no stored block is ever bet on twice and each harm bet's pairing gap is exactly one interval. The staleness family must in addition be *armed*: every rollback disarms it, each staleness bet consumes the armed flag, and only a subsequent fresh anchor measurement, which also refreshes its stored reference, re-arms it. Its reference at any bet is therefore the most recent anchor probe with no intervening rollback, each stored anchor realization is bet on at most once, and the post-cooloff reference gap is never bet. With the trial stream's consumption flags, the *one-bet-per-realization* invariant thus holds in all three families: harm by fresh pairing, gain by store-or-consume, staleness by the consumed arming flag. Because the adapter chooses its interval-t behavior before observing the gate's coins for any $t' \geq t$, and all certified statistics compare the probe against the already-realized pre-probe interval, no adaptation process can correlate its conduct with the coin outcomes. Predictable rates and predictable trial triggers change only what the sampled evidence represents (Lemma 4), never validity (Lemma 2). This division of labor is the mechanism behind Theorem 1 (the per-trial commit certificate) and Propositions 1–2 (certificate independence, validation certificate).

The applied filter ϕ_t. We write ϕ_t for the filter the supervisor actually applies at interval t: the candidate W_t on live intervals, the frozen candidate during cooloff (equal to the last rollback target, the anchor a_t, or 0 after a safe-side refuge rollback), and, on probe intervals, the anchor a_t or the fallback 0 according to the probe rules above.

Winsorized statistics and the certified functional. All tests operate on winsorized (clipped) block-level log-energy ratios,

$$Z_{j,b} = \mathrm{clip}\big(\log \ell^{(b)}(\text{cand}) - \log \ell^{(b)}(\text{ref}),\, \pm c\big),$$

where c is a design parameter (deployed: $c = 2$). The unit of betting evidence is the probe's *block-mean* $\bar{Z}_j = m^{-1}\sum_{b=1}^{m} Z_{j,b} \in [-c, c]$: each family bets exactly once per sampled probe, on $\bar{Z}_j$ or a transform of it. The certified functional is this statistic itself: we define

$$\varsigma_t \;=\; m^{-1}\sum_{b=1}^{m} \mathrm{clip}\big(\log \ell_t^{(b)}(\phi_t) - \log \ell_t^{(b)}(a_t), \pm c\big),$$

the applied system's *block-mean* winsorized log-excess over the anchor chain at interval t (with ϕ_t the applied filter defined above, and $\varsigma_t = 0$ identically on intervals where $\phi_t = a_t$). This is the same functional the bets sample: each sampled $\bar{Z}_j$ is a paired estimate of ς_t at the pre-probe live interval, exact up to one step of Assumption 2's drift, and the logged ledger of Proposition 3 accumulates these sampled values. Absolute (unwinsorized) energy damage is controlled separately through the hard bound of Lemma 1. The trade-off in c is discussed in App. A.3.

Remark (interval-level implication under block balance). Write $r_b = \ell^{(b)}(\text{cand})/\ell^{(b)}(\text{ref})$ and $w_b = \ell^{(b)}(\text{ref})/\sum_{b'} \ell^{(b')}(\text{ref})$. The *interval-level* log-energy excess is $\log \sum_b w_b r_b$, which our bets do not directly measure. Under a *block-balance* assumption (within each interval the block log-ratios have spread $\max_b \log r_b - \min_b \log r_b \le \Delta_{\mathrm{blk}}$) we have, on any interval where no block statistic clips, $\log \sum_b w_b r_b \le \max_b \log r_b \le \varsigma_t + \Delta_{\mathrm{blk}}$, so the certified statements transfer to the interval-level winsorized excess at the cost of $+\Delta_{\mathrm{blk}}$ per interval. On clipping intervals the absolute caps (Lemma 1, breaker) govern. We do not claim the transfer without block balance, and the main-text certificates are stated for the block-mean statistic.

Mixture betting (the deployed rule) and realized variation. Each family's wealth is a *uniform mixture* of fixed-fraction wealths over a finite grid $\mathcal{H}_f = \{\eta_1 < \cdots < \eta_{Q_f}\}$ of size Q_f,

$$E_n^f \;=\; \sum_{i=1}^{Q_f} \frac{1}{Q_f} \prod_{j \le n} \big(1 + \eta_i\, Z'_j\big), \tag{6}$$

with Z'_j the family's per-probe statistic. Deployed: $\mathcal{H}_{\mathrm{harm}} = \mathcal{H}_{\mathrm{stale}} = \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$ ($Q = 6$) and $\mathcal{H}_{\mathrm{gain}} = \{0.02, 0.05, 0.1, 0.2, 0.35\}$ ($Q = 5$). A convex combination of nonnegative supermartingales is again a nonnegative supermartingale, so anytime validity carries over from the fixed-fraction case at no cost in the error budget (Lemma 2). The restart-at-floor rule is applied to the mixture wealth, and all components restart together, so restarts remain stopping times of the mixture's own filtration. The wealth factors are positive at every grid point: $\eta \le 0.2 \ll 1/(c + \gamma/2) = 0.444$ for the harm/staleness statistics (raw range c, recentred range $c + \gamma/2$) and $\eta \le 0.35 < 1/(c + \gamma) = 0.4$ for the gain statistic. The harm grid stops at 0.2 because the detection bound of Proposition A(b) needs $\eta_{\max}(c + \gamma/2) \le 0.68$, which the deployed grid satisfies ($0.2 \times 2.25 = 0.45$), while a point at $\eta = 0.4$ would violate it ($0.9 > 0.68$).

Proposition A (App. A.5) quantifies the trade-off. A single fixed fraction η tolerates persistent harm below a floor set by the realized second moment of the statistic, and that fraction has to be guessed in advance without knowing the harm level. Mixing over the grid costs $\log Q$ nats, matches the best grid point in hindsight, and moves the floor to that of the smallest grid point, $\approx 2\bar{v}\eta_{\min}$, about $40\times$ lower. It does not remove it. When floor restarts are rare, the ledger of Proposition 3 is threshold-dominated ($\tilde{O}(\sqrt{T})$ total) down to that floor. Two derived quantities recur below. The *realized quadratic variation* of a family's run,

$$V_n \;:=\; \sum_{j \le n} (Z'_j)^2 \;\le\; c^2\, n \quad \text{(harm and staleness families; } (c+\gamma)^2 n \text{ for gain)},$$

is accumulated online by the gate (the gate logs it as a diagnostic, and it is not a tuning knob), and replaces the range bound in Lemma 3. Measured on the benchmark, $V_n/n = 1.81$ against the range $c^2 = 4$ (App. B). The *fixed-fraction sensitivity floor* is $h^*(\eta) := \eta\,\bar{v}$ with $\bar{v}$ an upper bound on $\mathbb{E}[(Z'_j)^2 \mid \mathcal{F}_{j-}]$: the average winsorized excess a fraction-η bet tolerates without ever detecting. Bounding $\bar{v}$ by the range c^2 (or by the recentred range $(c + \gamma/2)^2$) gives a fixed rate. The realized value is $2.2\times$ smaller than c^2 ($3\times$ smaller than the recentred range), and the certificate reports it.

A.2 AMPLIFICATION BOUND

Lemma 1 (Amplification bound)**.** *Let $y_{\max}$ be the driver saturation level, σ_n^2 the sensor-noise power, $\|S\|_1$ the secondary path's ℓ_1 norm, and $n_I = MH_b$ samples per interval. Then for every reachable θ, deterministically,*

$$\ell_t(\theta) \;\le\; 2\,\ell_t(0) + 2n_I\|S\|_1^2 y_{\max}^2.$$

Further, for any $\delta_n \in (0,1)$, with probability at least $1-\delta_n$ simultaneously for all $t \le T$,

$$\ell_t(0) \;\ge\; \tfrac{1}{2}\,\sigma_n^2\,n_I, \qquad \textit{provided } T\exp(-n_I/16) \le \delta_n,$$

and on that event $\ell_t(\theta)/\ell_t(0) \le B_{\max} := 2 + 4\|S\|_1^2 y_{\max}^2/\sigma_n^2$ for all t, θ.

Proof. Per sample $e = d + S*\sigma(y) + n$. By $(u+v)^2 \le 2u^2 + 2v^2$ with $u = d+n$ and $v = S*\sigma(y)$, and Young's inequality $\|S*\sigma(y)\|_2 \le \|S\|_1\|\sigma(y)\|_2 \le \|S\|_1 y_{\max}\sqrt{n_I}$, the deterministic upper bound follows, with $\ell_t(0) = \sum(d+n)^2$. For the lower bound, condition on the disturbance path: $\ell_t(0) \ge \sum_i (d_i + n_i)^2$ where, given d, the sum stochastically dominates a noncentral χ^2. In the worst case ($d \equiv 0$ pointwise) it is $\sigma_n^2\chi^2_{n_I}$, and the standard lower tail $\mathbb{P}(\chi^2_n \le n/2) \le \exp(-n/16)$ plus a union over $t \le T$ gives the claim. With $n_I = 256$, $\exp(-16) \approx 1.1 \times 10^{-7}$ per interval, negligible at any experimental horizon. Dividing the two bounds yields $B_{\max}$. □

With the benchmark constants ($\|S\|_1 \le 4$, $y_{\max} = 2$, $\sigma_n^2 = 10^{-4}$), $B_{\max} \approx 2.6 \times 10^6$ ($\approx 64\,\text{dB}$). The hard cap is a coarse physical backstop. The operational damage control during excursions comes from the circuit breaker, which restricts the candidate's noise-normalized error to $e^{c_{\text{brk}}}$ ($\approx 6.5\,\text{dB}$) above the anchor's recent level within $p = 2$ intervals. We report both constants alongside the certificates. The bound is deliberately *one-sided*: only amplification over passive needs a cap (a filter may of course be far *better* than passive), and no lower bound on $\ell_t(\theta)/\ell_t(0)$ is used anywhere. The tests handle both tails by winsorization.

A.3 WINSORIZATION: WHAT IS AND IS NOT CERTIFIED

Winsorization at $c < c_B := \log B_{\max}$ means the tests certify the *winsorized* mean of log-energy ratios. (i) Harm and staleness detections remain valid for winsorized harm, and any per-interval excess beyond e^c is invisible to the tests but still bounded absolutely by Lemma 1 and operationally by the breaker. (ii) The gain test certifies γ-improvement of the winsorized mean: a candidate whose loss ratio has mass beyond e^c on rare intervals can in principle be committed while being worse in raw energy. The probability of such tail mass is itself limited by the breaker (persistent $> c_{\text{brk}}$ excursions trigger rollback) but is not certified. Choosing $c \ge c_B$ recovers raw-energy certificates at the cost of a much less sensitive bet (the second moment $\bar{v}$ that sets the floor $h^*(\eta) = \eta\bar{v}$ grows quadratically in c). Deployed, the realized second moment is $\bar{v} = 1.81$ at $c = 2$, so a single fraction $\eta = 0.05$ tolerates ≈ 0.09 nats and the mixture $\approx 2\bar{v}\eta_{\min} \approx 0.02$. A larger c therefore buys a looser raw-energy statement at the cost of sensitivity. All main-paper claims are for the deployed $c = 2$ and are statements about winsorized log-loss plus the absolute caps.

A.4 VALIDITY OF THE RESTARTED BETTING FAMILIES

Lemma 2 (Validity, including mixtures)**.** *Fix a family with one statistic Z'_j per sampled probe, $|Z'_j| \le c + \gamma$, and null $H_0 : \mathbb{E}[Z'_j \mid \mathcal{F}_{j-}] \le 0$ (conditional on everything before that probe). For any η with $1 + \eta Z'_j > 0$ almost surely (harm/staleness: $\eta < 1/c = 0.5$, gain: $\eta < 1/(c+\gamma) = 0.4$, both satisfied at every deployed grid point) the wealth $E = \prod_j (1 + \eta Z'_j)$ is a nonnegative supermartingale with initial value 1. Consequently so is the mixture $E = \sum_i w_i \prod_j (1 + \eta_i Z'_j)$ of Eq. (6) for any fixed weights $w_i \ge 0$, $\sum_i w_i = 1$ (a convex combination of nonnegative supermartingales), and by Ville's inequality $\mathbb{P}(\sup E \ge 1/\delta_k) \le \delta_k$ for the mixture just as for a single component, so the grid consumes none of the error budget. Restart times (floor hits and event resets) are stopping times of the mixture (all components restart together), and by optional stopping, the process restarted with wealth 1 at a stopping time is again a nonnegative supermartingale conditionally on the stopping σ-algebra. With spending $\delta_k = \frac{\delta}{3} \cdot \frac{6}{\pi^2(k+1)^2}$ over restarts k, each family errs anywhere in time with probability at most $\delta/3$, and jointly at most δ. Serial correlation across probes is immaterial, since only the conditional-mean property is used. The selection of bet times does not matter either:*

any $\mathcal{F}$-predictable selection (scheduled coins at the predictable rate $1/\rho_t$, the fresh-pairing and arming eligibility flags, and the gain family's trial-probe schedule with its one-bet-per-realization consumption flags, both $\mathcal{F}_{t-1}$-measurable) preserves the supermartingale property, because the per-bet inequality $\mathbb{E}[1 + \eta Z'_j \mid \mathcal{F}_{j-}] \leq 1$ is conditional on everything before the bet. Predictable bet-time selection affects only what the sampled evidence represents (Lemma 4), never validity.

For the gain family the statistic is $Z'_j = -\bar{Z}^{\mathrm{tr}}_j - \gamma$ with $\bar{Z}^{\mathrm{tr}}_j$ the trial's block-mean pairing of the two gate-stored objects: at an anchor-side trial probe, the stored W_{tr} blocks against the probe's anchor blocks. At a snapshot-side trial probe, the probe's W_{tr} blocks against the stored anchor blocks (Theorem 1). Since $|\bar{Z}^{\mathrm{tr}}_j| \leq c$, $Z'_j \in [-(c+\gamma), c-\gamma]$ and the range condition holds. Apart from the transform and the betting grid the three families are identical, the harm and staleness families reading the certified stream and the gain family the trial stream.

Pairing bias. The harm and staleness families use one-sided pre-probe pairing under the fresh-pairing flag: a bet at probe t pairs the probe's blocks against the candidate blocks of interval $t-1$, realized before the probe coin was drawn. Under these families' nulls the statistic's conditional mean is therefore shifted by at most *one* step of conditional drift, κ (Assumption 2 applied once). No multi-step telescoping and no gap-conditioning martingale correction is needed: the machinery that pairing gaps $g_j > 1$ would require is unnecessary *by construction*, as is any multiplicity charge for consecutive probes re-betting stored blocks. The staleness family, whose reference refreshes only at anchor probes (and is bet on at most once: each bet consumes the arming flag), still carries multi-interval reference gaps. Those gaps enter no counting theorem (App. A.12). The margin γ_s is sized against their measured typical drift, reported in App. A.10. The gain family's trial bets carry in-trial gaps (between the betting probe and its stored opposite-side measurement), charged as κ_{tr} in Theorem 1. The commit condition must clear that transfer cost ($\gamma > \kappa_{\mathrm{tr}}$ plus margin). A symmetric (pre+post) pairing would cancel drift to first order, but the post-probe side is produced after the process has seen that a probe occurred, and an adversary can shape it. Only the pre-probe side is fixed before the coin is drawn, so we pair on that side alone and charge drift at its measured level. Worst-cell window p95 is 0.070 nats, with the $\leq 0.7\%$ exceedance ν-budgeted. Window medians (up to 0.019) are reported for context. No measured window in any suite or filter state reaches the $\gamma/2$ margin (App. A.10).

A.5 Proposition A: blind levels and calibration of deployment monitors

Here we isolate the statistical question from the ANC setting: every monitor below sees the same stream of winsorized paired statistics, and we ask what each can conclude from it. Two caveats come first. A contrast of "threshold versus sequential test" would be a strawman, since Page's CUSUM (Page, 1954) and its optimality theory (Lorden, 1971; Moustakides, 1986) are sequential, and adaptive or mixture charts are known in the quickest-change-detection literature (Sparks, 2000; Lai, 1998; Basseville & Nikiforov, 1993; Veeravalli & Banerjee, 2014). We also do not claim an asymptotic separation between "fixed-sensitivity" monitors and our mixture, because such a claim is not tuning-symmetric. A CUSUM whose reference is tuned to the horizon, $k \asymp \sqrt{\bar{v}\log(T/\delta)/T}$, attains the same $\tilde{O}(\sqrt{T})$ worst-case harm as a horizon-extended grid, and our own deployed grid, being finite, also retains a blind level of its own, which is where the measured worst case in fact sits. What we do claim is *(i)* a blind-level accounting under the cumulative-harm functional that covers Shewhart, CUSUM, EWMA, single-fraction betting *and the deployed mixture* in one framework, *(ii)* a detection-speed bound for the mixture above its floor, and *(iii)* a calibration asymmetry: at matched validity the e-process threshold is analytic, distribution-free and anytime-valid, while the charts' thresholds must be calibrated under the null law, which on-device drifts per user, fit and session, and on this statistic drives the calibrated Shewhart threshold to the clip, where it can never fire.

Definition 1 (Monitoring problem)**.** *A* monitor *observes winsorized statistics $Z_1, Z_2, \cdots \in [-c, c]$, one per sampled probe, and outputs a stopping time N (the intervention, in EGA the rollback). An* adversary *is any process generating the Z_j adaptively. Write $h_j := \mathbb{E}[Z_j \mid \mathcal{F}_{j-1}]$ for the conditional harm level. The* cumulative winsorized harm through T probes *is*

$$\mathrm{Harm}(T) \;:=\; \sum_{j \leq \min(N,T)} h_j,$$

harm accruing until the monitor intervenes (after an intervention the system runs the anchor, and the certified functional is 0 *by construction until adaptation resumes). A monitor is* δ-valid at horizon T *if* $\mathbb{P}(N \le T) \le \delta$ *against every null adversary (*$h_j \le 0$ *for all* j*).*

Definition 2 (Monitors and blind levels)**.** *A monitor has* blind level $\vartheta > 0$ *if it fires when a scalar statistic* S_j*, driven by the observations through parameters chosen before deployment, crosses a threshold, and* S *has non-positive conditional drift under every adversary holding* $h_j \le h < \vartheta$ *(for the memoryless monitor, whose statistic has no memory, read this as: the per-step firing probability stays at its null order). The five instances we compare:*

monitor	*statistic* S_j	*fires when*	*blind level* ϑ
Shewhart $\mathcal{M}_{\tau,p}$	Z_j *(run of* p*)*	$Z_{j-p+1}, \dots, Z_j > \tau$	τ
CUSUM (Page, 1954)	$\max(0, S_{j-1} + Z_j - k)$	$S_j > b$	k
EWMA	$(1-\alpha)S_{j-1} + \alpha Z_j$	$S_j > \tau$	τ
fixed-fraction e-process	$\sum_{i \le j} \log(1 + \eta Z_i)$	$S_j > A$	$b_- \eta v_{\mathrm{lo}}$
mixture e-process (deployed)	$\log \frac{1}{Q} \sum_\eta e^{\Lambda_j^\eta}$	$S_j > A_k$	$b_- \eta_{\min} v_{\mathrm{lo}}$

Here $\bar{v}$ *is the statistic's (average) second moment,* $v_{\mathrm{lo}} \le \bar{v}$ *a measured lower quantile of it (defined below the proof of this proposition.* $\bar{v}$ *upper-bounds* $\mathbb{E}[Z_j^2 \mid \mathcal{F}]$ *in part (b)'s hypothesis, so the two blind-level rows below use* v_{lo}*, the direction a blind-level argument actually needs), and* $b_- = 0.27$ *the constant in* $\log(1+x) \le x - b_- x^2$ *on the bet's range (not to be confused with the lower-bound constant* b_η *of Lemma 3). Rows one to three are the deployed divergence watchdog (*$\tau = c_{\mathrm{brk}}$*,* $p = 2$*) and the two standard quickest-change-detection charts (Basseville & Nikiforov, 1993; Veeravalli & Banerjee, 2014). Rows four and five are our own machinery without and with the mixture. A finite grid bottoms out at its smallest fraction, so mixing shrinks the blind level by* $\eta_{\max}/\eta_{\min} = 40\times$ *relative to the largest single fraction but does not remove it. The accounting indicts a* design pattern*, not a rival community, and it includes us.*

Write F for the fluctuation law of the statistic (mean zero, supported so that Z_j stays in $[-c, c]$), $q_F(u) := \mathbb{P}_F(\xi > u)$ its upper tail, F^{-1} its quantile function, and $\bar{v} := \mathbb{E}_F[\xi^2]$. No Gaussianity is assumed. All numerical instantiations below use the empirical law of Z logged by the harm family during real gated episodes (rms 1.26 nats, with an atom at the winsorization $\pm c$), which is what the corrected experiment of App. C.5 samples. Where a closed form is wanted we state the Gaussian-increment idealization separately and label it as such.

A lower bound for the blind-level claims. Part (b) below assumes $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \le \bar{v}$, an upper bound, and $\bar{v}$ (the realized average second moment, App. C.5) is a natural estimate of it. The blind-level claims (a4)–(a5) need the opposite direction: $\mathbb{E}[\log(1+\eta Z_j) \mid \mathcal{F}_{j-1}] \le \eta h_j - b_- \eta^2 \mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}]$ is only enlarged (weakened toward 0, which is what a blind-level proof needs) by replacing $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}]$ with a constant when that constant underbounds the true conditional second moment, i.e. we need $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \ge v_{\mathrm{lo}}$ for some v_{lo}, not $\le \bar{v}$. Hence (a4)–(a5) below use a separately measured $v_{\mathrm{lo}} \le \bar{v}$ in place of $\bar{v}$. We estimate it directly from the harm family's own logged draws (`gate.zbar_log`): partition the pooled, time-ordered sequence of block-means $\bar{Z}_j$ into non-overlapping windows of 32 probes, form the per-window sample second moment $\widehat{v}_w := \frac{1}{32} \sum_{j \in w} \bar{Z}_j^2$ as a windowed estimate of the local conditional second moment, and set v_{lo} to its 10th percentile over windows (`vlo_check.py`, `results/vlo.json`: $v_{\mathrm{lo}} = 1.596$ over 449 windows of 32 probes each, median window value 2.152). By its definition v_{lo} underbounds the realized conditional second moment on $\approx 90\%$ of windows, not all of them. Like the windowed drift constants of App. A.10, it is a measured quantity, not a proven bound. We therefore state (a4)–(a5) as holding on the $(1-\nu_v)$-fraction of windows with $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \ge v_{\mathrm{lo}}$, $\nu_v \approx 0.10$ by the choice of quantile. Outside that fraction we make no blind-level claim (the monitor may in fact detect faster there, since a smaller realized variance than v_{lo} only helps the drift argument. ν_v bounds how often the argument is silent, not how often it is wrong).

Proposition A (full statement). *Fix* c*, a horizon* T *(in probes), and a fluctuation law* F*. Throughout part (a) the adversary is the constant-harm one,* $Z_j = \mathrm{clip}(h + \xi_j)$ *with* $\xi_j \sim F$ *i.i.d. and* $\mathbb{E}[Z_j] \le h$*, and* N *is the monitor's firing time.*

(a) Every monitor of Definition 2, the deployed mixture among them, has a blind spot.

(a1) Walk-type statistics (CUSUM, and any monitor of the form $S_j = \max(0, S_{j-1} + Z_j - \vartheta)$ firing at $S_j > b$). *For $h < \vartheta$ let $\theta_h > 0$ solve $\mathbb{E}[e^{\theta(Z-\vartheta)}] = 1$ (it exists and is unique whenever $\mathbb{E}[Z] < \vartheta$ and $\mathbb{P}(Z > \vartheta) > 0$, and θ_h is strictly decreasing in h). Then*

$$\mathbb{P}(N \le T) \;\le\; T\, e^{-\theta_h b}, \qquad \mathrm{Harm}(T) \;\ge\; h\,T\,\big(1 - T e^{-\theta_h b}\big). \tag{7}$$

Suppose in addition the threshold is validity-calibrated *in the classical sense: δ-validity at horizon T requires an average run length to false alarm of order T/δ, hence*

$$b \;\ge\; \theta_0^{-1} \log\big(C_F\, T/\delta\big) \tag{8}$$

for a law-dependent constant $C_F > 0$ (the ARL–threshold relation; Lorden, 1971; Moustakides, 1986; Basseville & Nikiforov, 1993). Then $\mathrm{Harm}(T) \ge \frac{3}{4} hT$ for every $h < \vartheta$ with $\theta_h/\theta_0 \ge \log(4T)/\log(C_F T/\delta)$. In the Gaussian-increment idealization, where $\theta_h = 2(\vartheta - h)/\bar{v}$ exactly and hence $\theta_h/\theta_0 = 1 - h/\vartheta$, this holds for all $h \le \vartheta\, \log(C_F/(4\delta))/\log(C_F T/\delta)$, so

$$\sup_h \mathrm{Harm}(T) \;\ge\; \tfrac{3}{4}\,\vartheta\, T \cdot \frac{\log(C_F/4\delta)}{\log(C_F T/\delta)} \;=\; \Omega\Big(\frac{\vartheta\, T \log(1/\delta)}{\log T}\Big) \;=\; T^{1-o(1)}. \tag{9}$$

Intuitively, $S_j = \max(0, S_{j-1} + Z_j - k)$ drifts downward whenever $\mathbb{E}[Z_j] < k$, so harm below the reference never accumulates. This is the blind spot of a raw threshold moved down to k, and a smaller k does not help because $\vartheta = k$ also sets the false-alarm rate and (8) *then forces b up.*

(a2) Averaging statistics (EWMA at level τ, smoothing α). *For $h < \tau$,*

$$\mathbb{P}(N \le T) \;\le\; T \exp\Big(-\frac{(\tau - h)^2(2-\alpha)}{2c^2\alpha}\Big), \tag{10}$$

so with τ validity-calibrated, i.e. $\tau^2(2-\alpha)/(2c^2\alpha) \ge \log(T/\delta)$, the same optimization gives $\sup_h \mathrm{Harm}(T) = \Omega(\tau T \log(1/\delta)/\log T)$: blind level $\vartheta = \tau$, identical shape.

(a3) Memoryless statistics (Shewhart $\mathcal{M}_{\tau,p}$). *For every $h \in (0, \tau)$,*

$$\mathrm{Harm}_{\mathcal{M}_{\tau,p}}(T) \;\ge\; \tfrac{3}{4}\, h \cdot \min\Big(T,\; \tfrac{1}{2}\, q_F(\tau - h)^{-p}\Big), \tag{11}$$

and δ-validity at horizon T forces

$$\tau \;\ge\; \tau_{\min}(T) \;:=\; F^{-1}\Big(1 - \big(\tfrac{p}{T} \log \tfrac{1}{1-\delta}\big)^{1/p}\Big), \tag{12}$$

so taking $h_T := \tau - F^{-1}(1 - (2T)^{-1/p}) > 0$ gives $\mathrm{Harm} \ge \frac{3}{4} h_T T = \Omega(T)$. On a winsorized statistic this instance degenerates completely: if F has an atom of mass $a > 0$ at the clip, then $\tau_{\min}(T) = c$ as soon as $Ta^p \ge p \log \frac{1}{1-\delta}$, in which case $\mathbb{P}(Z_j > \tau) = 0$, the monitor never *fires, and $\mathrm{Harm}(T) = hT$ for* every *$h \le c$, blind on the whole range, $\Theta(T)$ with the maximal constant. This is not a pathology of the analysis: it is what the calibration in App. C.5 returns ($\tau = 2.00 = c$).*

(a4) A fixed betting fraction is in the class, on the $(1 - \nu_v)$-fraction of windows with $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \ge v_{\mathrm{lo}}$. *Let $b_- := 0.27$, so that $\log(1+x) \le x - b_- x^2$ on the bet's range $|x| \le 0.68$. If the harm family is run at a single fraction η with threshold A, its log-wealth is the walk of (a1) with increment $g(z) = \log(1 + \eta z)$, whose mean is non-positive for every $h \le \vartheta := b_- \eta v_{\mathrm{lo}}$, on that fraction of windows. Hence (a1) applies unchanged (restricted to those windows), $\mathbb{P}(N \le T) \le T e^{-\theta_h A}$ with $\theta_h > 0$ solving $\mathbb{E}[e^{\theta g(Z)}] = 1$, and the monitor is blind below $\vartheta = b_- \eta v_{\mathrm{lo}}$: harm $T^{1-o(1)}$.*

(a5) So is the deployed mixture, on the same fraction of windows. *For $h \le \vartheta := b_- \eta_{\min} v_{\mathrm{lo}}$ every component's log-wealth is a non-positive-drift walk simultaneously, and a union over the Q components gives $\mathbb{P}(N \le T) \le Q\, T\, e^{-\theta_h^{\min} A_k}$ with $\theta_h^{\min}$ the smallest component Lundberg exponent: the mixture is blind below ϑ. Deployed ($\eta_{\min} = 0.005$, $v_{\mathrm{lo}} = 1.596$): $\vartheta \approx 0.002$ nats, and the residual linear branch of part (b) (which legitimately uses the upper bound $\bar{v} = 1.81$, not v_{lo}) is $2\bar{v}\eta_{\min} T \approx 0.018\, T$, a distinct, looser quantity (the sub-grid-floor term of the worst-case-*harm *bound, not the blind level itself. The two are never interchanged elsewhere in this paper), which is where the measured worst case ($0.0194\, T$, attained at the smallest tested level $h = 0.02$) in fact sits. Mixing shrinks the blind level $40\times$ relative to the largest single fraction. It does not remove it.*

(b) Upper bound for the mixture e-process. *Let the harm family run the mixture* (6) *over a grid $\mathcal{H} \subset (0, \eta_{\max}]$ with Q points, $\eta_{\max}(c + \gamma/2) \le 0.68$, and consecutive ratio at most ρ. Suppose an adversary holds $h_j \ge h > 0$ with conditional second moments $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \le \bar{v}$, and let*

$A' := A_k + \log Q$. *If the grid covers the optimal fraction,* $h/(2\bar{v}) \in [\eta_{\min}, \eta_{\max}]$ *with* $\eta_{\min} := \min \mathcal{H}$ *(levels above* $2\bar{v}\eta_{\max}$ *are detected at least as fast, with drift* $\geq \eta_{\max} h/2$*), then for any* $\delta' \in (0,1)$ *the mixture crosses its threshold within*

$$N(h) \;\leq\; \rho^2\, \frac{8\,\bar{v}\,A' \;+\; 648\log(1/\delta')}{h^2} \;=\; O\Big(\frac{\bar{v}\,\log(Q/(\delta_k\delta'))}{h^2}\Big) \quad \textit{sampled probes, w.p. } \geq 1-\delta', \tag{13}$$

whence $\mathrm{Harm}(T) \leq \min\{hT,\, hN(h)\}$ *and, maximizing over the adversary's choice of* h*,*

$$\sup_{h>0} \mathrm{Harm}(T) \;\leq\; \max\Big(2\bar{v}\,\eta_{\min} T,\; \sqrt{\rho^2\big(8\bar{v}A' + 648\log(1/\delta')\big)\,T}\Big) \;=\; O\big(\sqrt{\bar{v}\,T\log(QT/\delta)}\big), \tag{14}$$

the first branch being the residual floor of a grid truncated at $\eta_{\min}$*. The floor can only be removed with horizon information. Adding* $\eta \asymp \sqrt{A'/(\bar{v}T)}$ *to the grid costs* $\log Q \approx \log\log T$ *nats and lets the second branch bind up to* T*, but a CUSUM given the same information can tune its reference* $k \asymp \sqrt{\bar{v}\log(T/\delta)/T}$ *and reach the same* $\tilde{O}(\sqrt{T})$ *worst-case rate. The rate is therefore* not *the distinguishing claim.*

(c) Calibration at matched validity. *The mixture's threshold* $A_k = \log(1/\delta_k)$ *is analytic, distribution-free, and valid at every stopping time and horizon simultaneously (Ville). No property of* F *enters. For the charts,* δ*-validity pins the threshold to the null law: Shewhart needs* $\tau \geq \tau_{\min}(T)$ *of* (12) *(on a winsorized statistic with an atom at the clip this equals* c *at any realistic* T*, and the monitor can never fire), while the CUSUM and EWMA thresholds are roots of ARL-type relations* (8) *and* (10) *in law-dependent constants that we, with full offline access to the null pool, could only obtain by simulation (App. C.5:* $b = 18.7$*,* $\tau = 0.65$*). On a device the null law varies per user, fit and session and drifts within one: a simulation-calibrated threshold is exactly what cannot be shipped, and a miscalibrated one is either invalid (false alarms) or blind (inflated* ϑ*). This asymmetry, not an asymptotic rate, is the operational content of parts (a)–(b). The measured constant factors (*$3.7\times$*/*$3.8\times$*/*$41.2\times$ *at* $T = 20{,}000$*) quantify it at deployment horizons.*

Proof. (a1). The reflected recursion unrolls to $S_j = \max_{0\leq i\leq j} \sum_{l=i+1}^{j}(Z_l - \vartheta)$. Fix a start index i and put $M_j^{(i)} := \exp\big(\theta_h \sum_{l=i+1}^{j}(Z_l - \vartheta)\big)$ for $j > i$. By the definition of θ_h and independence, $\mathbb{E}[M_j^{(i)} \mid \mathcal{F}_{j-1}] = M_{j-1}^{(i)}\,\mathbb{E}[e^{\theta_h(Z_j-\vartheta)}] = M_{j-1}^{(i)}$, so $M^{(i)}$ is a nonnegative martingale started at 1 and Ville's inequality (Ville, 1939) gives $\mathbb{P}\big(\exists j > i : \sum_{l=i+1}^{j}(Z_l - \vartheta) > b\big) \leq e^{-\theta_h b}$. A union bound over the at most T start indices that can be active before time T gives $\mathbb{P}(N \leq T) \leq Te^{-\theta_h b}$, and $\mathrm{Harm}(T) = h\,\mathbb{E}[\min(N,T)] \geq hT\mathbb{P}(N > T)$ gives (7). Existence, uniqueness and monotonicity of θ_h are the standard Lundberg facts for a bounded increment with negative mean: $\theta \mapsto \log \mathbb{E}[e^{\theta(Z-\vartheta)}]$ is convex, vanishes at 0 with negative slope $\mathbb{E}[Z] - \vartheta < 0$, and tends to $+\infty$ since $\mathbb{P}(Z > \vartheta) > 0$. Raising h raises the function pointwise and therefore lowers its positive root. Combining with the calibration bound (8): if $\theta_h/\theta_0 \geq \log(4T)/\log(C_F T/\delta)$ then $\theta_h b \geq \log(4T)$, so $Te^{-\theta_h b} \leq \frac{1}{4}$ and $\mathrm{Harm} \geq \frac{3}{4}hT$. In the Gaussian idealization $\log \mathbb{E}[e^{\theta(Z-\vartheta)}] = \theta(h-\vartheta) + \theta^2\bar{v}/2$ has positive root $\theta_h = 2(\vartheta - h)/\bar{v}$, so $\theta_h/\theta_0 = 1 - h/\vartheta$ and the admissible levels are $h \leq \vartheta(1 - \log(4T)/\log(C_F T/\delta)) = \vartheta\log(C_F/4\delta)/\log(C_F T/\delta)$, which is (9).

(a2). $S_j = \alpha\sum_{l\leq j}(1-\alpha)^{j-l}Z_l$ is a weighted average of independent variables in $[-c, c]$ with weights summing to at most 1 and squared weights summing to at most $\alpha^2/(1-(1-\alpha)^2) = \alpha/(2-\alpha)$. Since $\mathbb{E}[S_j] \leq h$, Hoeffding's inequality gives $\mathbb{P}(S_j > \tau) \leq \exp(-(\tau-h)^2(2-\alpha)/(2c^2\alpha))$, and a union bound over $j \leq T$ gives (10). Optimizing as in (a1) (the calibrated τ makes the null exponent $\log(T/\delta)$, the harm bound needs exponent $\log(4T)$, and the two differ by a factor $1 - h/\tau$ in the square root) gives the stated $\Omega(\tau T/\log T)$.

(a3). Under the constant-harm adversary the exceedance indicators $1\{Z_j > \tau\}$ are i.i.d. Bernoulli(q) with $q := q_F(\tau - h)$, and N is the first completion of a run of p ones. A union bound over the at most t positions at which such a run can complete gives $\mathbb{P}(N \leq t) \leq tq^p$, so with $T' := \min(T, \frac{1}{2}q^{-p})$,

$$\mathbb{E}[\min(N,T)] \;\geq\; \int_0^{T'} \mathbb{P}(N > t)\,dt \;\geq\; \int_0^{T'} (1 - tq^p)\,dt \;=\; T'\big(1 - \tfrac{T'q^p}{2}\big) \;\geq\; \tfrac{3}{4}\,T',$$

and $\mathrm{Harm}(T) = h\,\mathbb{E}[\min(N,T)]$ gives (11). For $h = h_T$ we have $q_F(\tau - h_T) = (2T)^{-1/p}$, hence $\frac{1}{2}q^{-p} = T$ and $T' = T$. If $F = \delta_0$ then $q = 0$ for every $h < \tau$ and the monitor never fires. For (12),

partition $\{1,\dots,T\}$ into $\lfloor T/p \rfloor$ disjoint blocks of length p. Under the null adversary $Z_j = \xi_j$ the blocks are independent and each is all-exceedances with probability $q_F(\tau)^p$, so $\mathbb{P}(N \le T) \ge 1-(1-q_F(\tau)^p)^{\lfloor T/p \rfloor} \ge 1-\exp(-\lfloor T/p \rfloor q_F(\tau)^p)$. Requiring this to be at most δ forces $q_F(\tau)^p \le (p/T)\log\frac{1}{1-\delta}$, i.e. $\tau \ge \tau_{\min}(T)$. Positivity of the blind spot $\tau_{\min}(T) - F^{-1}(1-(2T)^{-1/p})$ follows since $(p/T)\log\frac{1}{1-\delta} < (2T)^{-1} \le (2T)^{-1/p}\cdot(2T)^{1-1/p}$ whenever $p\log\frac{1}{1-\delta} < \frac{1}{2}$ (deployed: $p=2$, $\delta = 0.05$, value 0.10), and F^{-1} is decreasing in its tail argument. For the winsorized degeneracy, $q_F(u) \ge a$ for every $u < c$ when F has an atom of mass a at the clip, so $q_F(\tau)^p \le (p/T)\log\frac{1}{1-\delta}$ is unsatisfiable for $\tau < c$ once $Ta^p \ge p\log\frac{1}{1-\delta}$. At $\tau = c$ the event $\{Z_j > c\}$ is empty because the statistic is winsorized, so $N = \infty$ and $\mathrm{Harm}(T) = hT$ identically.

(a4). On $|x| \le 0.68$ one has $\log(1+x) \le x - b_- x^2$ with $b_- = 0.27$ (the cubic remainder of $\log(1+x)$ is dominated there), so $\mathbb{E}[\log(1+\eta Z_j) \mid \mathcal{F}_{j-1}] \le \eta h_j - b_-\eta^2 \mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}]$. On the $(1-\nu_v)$-fraction of windows where $\mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \ge v_{\mathrm{lo}}$ (the direction v_{lo} is measured to support, unlike $\bar v$, which bounds this quantity from above in part (b)), this is $\le \eta\,(h - b_-\eta v_{\mathrm{lo}}) \le 0$ whenever $h \le \vartheta = b_-\eta v_{\mathrm{lo}}$. On those windows the log-wealth is then a random walk with non-positive mean and increments bounded by $|\log(1 \pm \eta c)|$, restarted at its floor as in (a1). The argument of (a1) applies with $g(z) = \log(1+\eta z)$ and $b = A$, restricted to the sub-filtration of windows satisfying the second-moment floor (a predictable restriction, so the martingale argument of (a1) is unaffected).

(b). Let η be the largest grid point with $\eta \le h/(2\bar v)$. It exists and satisfies $\eta \ge h/(2\rho\bar v)$ because the grid covers $h/(2\bar v)$ with consecutive ratio at most ρ. Write $X_j := \log(1+\eta Z_j)$. Since $\eta(c+\gamma/2) \le 0.68$, the elementary bound $\log(1+x) \ge x - x^2$ (valid for $x \ge -0.68$) gives

$$\mathbb{E}[X_j \mid \mathcal{F}_{j-1}] \;\ge\; \eta h_j - \eta^2 \mathbb{E}[Z_j^2 \mid \mathcal{F}_{j-1}] \;\ge\; \eta h - \eta^2 \bar v \;\ge\; \frac{h^2}{4\rho\bar v} \;=:\; D,$$

using $\eta \in [h/(2\rho\bar v), h/(2\bar v)]$. The increments satisfy $|X_j| \le |\log(1-\eta(c+\gamma/2))| \le 2\eta(c+\gamma/2) \le 4.5\,\eta \le 2.25\,h/\bar v =: M$ (using $|\log(1-x)| \le 2x$ for $x \le 0.68$). Centering, $\sum_{j\le n}(X_j - \mathbb{E}[X_j \mid \mathcal{F}_{j-1}])$ is a martingale whose increments lie in an interval of length $2M$ (as $X_j \in [-M, M]$), so by Azuma–Hoeffding, with probability $\ge 1-\delta'$, $\sum_{j\le n} X_j \ge nD - M\sqrt{2n\log(1/\delta')}$. Taking $n = 2A'/D + 8M^2\log(1/\delta')/D^2$ and applying $2\sqrt{ab} \le a+b$ yields $\sum_{j\le n} X_j \ge A'$, i.e. the η-component's wealth exceeds $e^{A'} = Qe^{A_k}$ and the uniform mixture exceeds $e^{A_k} = 1/\delta_k$. Substituting $D \ge h^2/(4\rho\bar v)$ and $M \le 2.25h/\bar v$: $2A'/D \le 8\rho\bar v A'/h^2$ and $8M^2\log(1/\delta')/D^2 \le 648\,\rho^2\log(1/\delta')/h^2$, which is (13) (the constants are not optimized: a Freedman bound with predictable variance $\le C\eta^2\bar v$ would replace 648 by an $O(\bar v)$ constant). For (14), harm accrues at rate h until detection, so $\mathrm{Harm}(T) \le h\min(T, N(h))$. The two branches hT and C_A/h (with C_A the numerator of (13)) cross at $h = \sqrt{C_A/T}$, and for $h < 2\bar v\eta_{\min}$ (where the grid can no longer supply $\eta \le h/(2\bar v)$) the harm is at most $2\bar v\eta_{\min}T$ trivially.

(a5). On the same $(1-\nu_v)$-fraction of windows as (a4), for $h \le b_-\eta_{\min}v_{\mathrm{lo}}$ and every grid point $\eta_i \ge \eta_{\min}$, $\mathbb{E}[\log(1+\eta_i Z_j) \mid \mathcal{F}_{j-1}] \le \eta_i h - b_-\eta_i^2 v_{\mathrm{lo}} = \eta_i(h - b_-\eta_i v_{\mathrm{lo}}) \le 0$, so each component log-wealth Λ^{η_i} is a walk with non-positive drift there. The mixture exceeds A_k only if $\frac{1}{Q}\sum_i e^{\Lambda^{\eta_i}} \ge e^{A_k}$, which forces $\max_i \Lambda^{\eta_i} \ge A_k$. A union over the Q components and the argument of (a1) applied to each give $\mathbb{P}(N \le T) \le Q\,T\,e^{-\theta_h^{\min}A_k}$.

(c). The Shewhart degeneracy is the atom argument of (a3). The remaining claims are properties of the thresholds themselves, quantified by the calibration outputs of App. C.5. □

Remark (restarts). Part (b) is stated for a run without a floor restart. Under sustained harm the log-wealth has positive drift D and increments bounded by M, so by the same Azuma bound the probability that a run of length n ever falls below its floor decays as $\exp(-\Theta(nD^2/M^2))$. The standard e-detector argument (Shin et al., 2024) then inflates $N(h)$ by a constant factor and replaces A_k by $A_{k'}$ at the (logarithmically growing) restart index. The synthetic comparison behind the deployed numbers runs the deployed restart rule, so the measured detection times already include this effect.

Deployed instantiation. The blind levels of Def. 2 take concrete values on the benchmark. Calibrating each monitor under the empirical null to a common $\delta/3$ false-alarm budget over $T = 20{,}000$ probes (43 minutes, App. C.5, `results/separation_v2.json`) yields $\vartheta = 2.00$ for the Shewhart

chart (equal to the winsorization $c = 2$, so case (a3) is in its degenerate branch and the monitor is blind on the entire range), $\vartheta = k = 0.25$ at threshold $b = 18.7$ for CUSUM, and $\vartheta = \tau = 0.65$ at $\alpha = 0.05$ for EWMA. The mixture's threshold is the analytic $\log(1/\delta_k)$ under any null law and needs no calibration. Measured worst-case harms over $h \in [0.02, 0.8]$ are $0.0194\,T$ (mixture), $0.0720\,T$ (CUSUM), $0.0733\,T$ (EWMA) and $0.8000\,T$ (Shewhart, i.e. hT at the largest tested h, as (a3) forces), and the growth exponents across $T = 4000 \to 20000$ are 0.53, 0.86, 0.73 and 1.00. The exponents are measured rather than derived: Shewhart's exact 1.00 follows from (a3), and CUSUM's and EWMA's lie near the $1 - \Theta(1/\log T) \approx 0.92$ that (a1)–(a2) suggest between these horizons. The mixture's 0.53 is pre-asymptotic, since its worst case is attained at the grid boundary $h = 0.02$ and is floor-limited by (a5). We neither optimize the constants in (9) and (13) nor fit them to the data.

Remark (the effect of mixing). The blind level of a fixed fraction is $b_- \eta v_{\mathrm{lo}}$ (on the $(1 - \nu_v)$-fraction of windows of the preceding remark) with $v_{\mathrm{lo}} \le \bar{v}$ a measured lower quantile of the realized second moment, and not $\eta(c + \gamma/2)^2$, which bounds $\mathbb{E}[Z_j^2]$ by the range and is an artifact of that bound (on this benchmark $v_{\mathrm{lo}} = 1.596$, $\bar{v} \approx 1.81$, against $c^2 = 4$). A fixed $\eta = 0.05$ therefore still detects persistent $h \gg 0.05\, b_- v_{\mathrm{lo}}$. Part (a4) says that it has a blind spot of the same kind as CUSUM's, and that shrinking η shrinks the blind spot and the detection speed at large h together. Mixing lowers the blind level $40\times$ ($b_- \eta_{\min} v_{\mathrm{lo}} \approx 0.002$), so the $\tilde{O}(\sqrt{T})$ branch of part (b) holds down to the grid floor. Mixing also means $\bar{v}$ (the scale of h) need not be known at design time, since the min-over-grid bound of Lemma 3 is an empirical-Bernstein statement that adapts to the realized variation, and the threshold stays analytic and anytime-valid rather than null-calibrated, the deployable difference (part (c)). All of this costs $\log Q$ nats once. What it does not buy: detection below the floor (nothing detects $h \le 0.05$ within 43 minutes, App. C.5), or an asymptotic rate advantage over a horizon-tuned chart.

A.6 Certificate independence of the virtual-probe channel

Proposition 2 (full statement). *Fix any calibration model $\hat{S}$ (with arbitrary, even adversarially constructed, error) and any adaptation process π. Then every certificate of this paper holds unchanged, with the same constants and the same probabilities, for every realization of the virtual-probe channel: the validity of the betting families (Lemma 2), the logged harm ledger (Proposition 3), the per-trial commit certificate (Theorem 1), the validation certificate (Proposition 1), and the trial phase of Theorem 2 (whose proposal phase is explicitly model-dependent).*

Proof. The channel consists of (i) the trust state driving $K_t \in \{K, K_{\mathrm{slow}}\}$ (with the alert window pinning the fast rate for a_{al} probes after any rollback-like event, trust intact), (ii) the trial proposals, which freeze W_{tr} and add extra trial draws at rate $1/P_{\mathrm{tr}}$ beside the never-paused base coins, and (iii) the virtual-harm reflex. All three are deterministic functionals of past measurements, past coins and $\hat{S}$, hence $\mathcal{F}_{t-1}$-measurable at the moment they act. They read no future coin, and virtual statistics never enter a certified wealth. We go through the certificates in turn.

Validity and the ledger. The harm and staleness families bet only on certified base coins, which run at every interval, even during trial windows. The conditional law of G_t given $\mathcal{F}_{t-1}$ is exactly Bernoulli$(1/\rho_t)$ with ρ_t predictable, and Lemma 2 uses only the per-bet conditional inequality, so each family's wealth remains a nonnegative supermartingale under its null for any $\hat{S}$. The ledger of Proposition 3 is pathwise over the sampled bets: its inequality holds on every realization regardless of which intervals the trust state caused to be sampled. The model changes only which sample is realized: the probe rate, hence the coverage and the per-tier rates that Lemma 4 reports, all visible in the logs, while the inequality and its error budget are unaffected.

Commits. The gain family bets at trial probes. These are predictable times, covered by the same argument (Theorem 1), and each trial certifies a gate-stored frozen object regardless of what $\hat{S}$ proposed. A wrong model can only waste trials (each failed trial costing its own probes and cooldown), never promote an uncertified object. Reflex rollbacks are safe retreats that cost performance, not validity.

No step of any kept proof references the accuracy of $\hat{S}$. The model changes which intervals are probed and how often, a performance effect, and nothing else. Conversely, the performance cost of a wrong model is bounded by construction: proposals waste at most one trial each (with its cooldown),

reflexes only retreat, and distrust restores the full-rate schedule, whose certificates never depended on the model in the first place. □

A.7 THE VALIDATION CERTIFICATE (CERTIFIED TRUST)

In the headline configuration (reflexes + virtual probes), which Sec. 5 states in full, the validation family is the only e-process still acting. It sets the probe rate and hence the supervision tax, and it gates the model's proposals, so it is the certificate that configuration carries. We restate its guarantee formally (main-text statement: Prop. 1).

Proposition 1 (full statement). *At each fresh anchor probe j carrying a stored virtual prediction, let g_j be the* validation gap*: the virtual prediction of the anchor's mean log-power minus its real measured value. Define the winsorized deviation statistics $Z_j^u := \mathrm{clip}(g_j - \gamma_v, \pm c)$ and $Z_j^d := \mathrm{clip}(-g_j - \gamma_v, \pm c)$, with validation margin $\gamma_v = 0.25$ nats. The validation family runs two mixture wealths (Eq.* (6)*, grid $\mathcal{H}_{\mathrm{harm}}$), one per side, restarted together at their joint floor with a shared restart index, and* distrust *fires when either crosses its threshold.*

(a) (Anytime-valid calibration test.) *Say the virtual-probe channel is γ_v-*calibrated *at probe j if $\mathbb{E}[Z_j^u \mid \mathcal{F}_{j-}] \le 0$ and $\mathbb{E}[Z_j^d \mid \mathcal{F}_{j-}] \le 0$ (implied by $|\mathbb{E}[g_j \mid \mathcal{F}_{j-}]| \le \gamma_v$ whenever clipping is inactive). Under the null that the channel is γ_v-calibrated at every bet, each side's wealth is a nonnegative supermartingale at every grid point, and with the spending schedule the two-sided test errs anywhere in time with probability at most $2\delta/3$ (a budget independent of the three filter-level families' shared δ, see the bookkeeping remark of Sec. 5). Every distrust event is therefore an anytime-valid rejection of the channel's calibration. The certificate concerns the evidence channel, not any filter.*

(b) (All model influence is gated or safe.) *The calibration model touches the system only through $\mathcal{F}_{t-1}$-measurable decisions: the trust-gated probe rate K_t (thinned only after six validated probes, restored by distrust and demotion, with the alert window an $\hat{S}$-independent rule), trial proposals (which freeze a gate-stored object whose promotion requires real paired evidence, Theorem 1), and the certificate-free virtual-harm reflex, which is left ungated by trust (it runs in every configuration with virtual probes enabled) because a false rollback is safe. Proposition 2 therefore holds verbatim with trials and alert mode included: model error selects the realized tier and the trial occupancy, a performance effect, never validity.*

(c) (Role in the reflex configuration.) *With the filter-level families disabled, (a) is the certificate that ships: trust (and with it the thinned tax and every model-proposed action) is never extended except while the virtual channel withstands an anytime-valid calibration test at an analytic, distribution-free threshold.*

Proof. (a) is Lemma 2 applied to Z_j^u and Z_j^d. The bet times are $\mathcal{F}$-predictable: whether a stored prediction exists, whether the previous interval ran the candidate (fresh pairing), and the probe target are all determined by $\mathcal{F}_{j-}$ and the gate's own coins. The stored prediction is $\mathcal{F}_{j-}$-measurable, so each statistic is a fixed measurable transform of the probe's measurement, and the per-bet conditional inequality is the null. The two sides restart together at the joint floor (a stopping time of the pair) and share one restart index. A union over the two sides doubles the family budget to $2\delta/3$. (This budget is accounted separately from the δ of the filter-level certificates: a validation error costs trust, i.e. performance, and never enters the filter-level certificates.) (b) is the proof of Proposition 2 read as an exhaustive list: those are the only code paths by which $\hat{S}$ reaches the system, no virtual statistic enters a certified wealth, and every gated path is supervised by the trust state, which transitions only on validation crossings, the validated-probe count, and demotions. The reflex is the one ungated path, and its sole action is a rollback. (c) is a configuration fact of the implementation: with the test families disabled, no proposal or trial ever opens (so no gain bet is placed) and no filter-level wealth can trigger a rollback, demotion or commit (the harm and staleness wealths are maintained only as logged diagnostics), while the validation family, the virtual-harm reflex, the breaker and the backoff all still run. □

Remark (scope of the certificate). The proposition certifies the *channel*, meaning virtual predictions of anchor error agree with real anchor measurements to within γ_v in winsorized conditional mean, not any filter-level property. The filter-level budget remains the optional tier below (Theorem 1,

Proposition 3, App. A.11). Its detection power is the measured validation SNR: under the polarity fault the gap is $\approx 1.6 \pm 1.1$ nats (a few probes to distrust), post-fault re-seats $\approx 0.6 \pm 1.2$. Benign re-seats are the weak case (near-zero mean gap), covered by the K_{slow} certificate floor and by the rule that commits require real trial evidence regardless of trust.

The certified tier (optional apparatus). Every result above (validity, Proposition A, certificate independence, and the validation certificate of the reflex tier) holds whether or not the harm/gain/staleness e-tests are enabled. So do the two generic lemmas that open the next subsection (predictable-rate sampling and the wealth-regret bound), which are stated once and reused below. What is specific to the optional certified tier (Sec. 4) begins with the logged harm ledger (Prop. 3) itself and continues through the per-trial commit certificate, what it does not establish, and the trial/adoption latency analysis (App. A.13). These are the only consumers of Assumption 2's drift constant and of the staleness-suspicion occupancy fraction that Prop. 3 excludes. Neither code path is open in the reflex tier, so Prop. 1 does not depend on either constant.

A.8 THE LOGGED HARM LEDGER: PATHWISE WEALTH-REGRET

Lemma 3 (Winsorized harm regret)**.** *Consider any contiguous run of the harm family between two restarts, with threshold $A_k = \log(1/\delta_k)$. Explicitly, since $\delta_k = \frac{\delta}{3} \cdot \frac{6}{\pi^2(k+1)^2} = \frac{2\delta}{\pi^2(k+1)^2}$,*

$$A_k \;=\; \log(1/\delta_k) \;=\; \log\frac{\pi^2(k+1)^2}{2\delta} \qquad \textit{(the factor } 6/\pi^2 \textit{ makes } \textstyle\sum_{k\geq 0}\delta_k = \delta/3 \textit{ exactly).}$$

Let $Z_1, \ldots, Z_n$ be the sampled per-probe block-mean statistics during the run ($|Z_i| \leq c$ for the harm and staleness families) and let $V_n := \sum_{i\leq n} Z_i^2 \leq c^2 n$ be their realized quadratic variation*. Write $\mathcal{H}$ for the family's betting grid ($Q := |\mathcal{H}|$), $A'_k := A_k + \log Q$ for the mixture-inflated threshold, and, for $\eta \in \mathcal{H}$, $b_\eta := 1$ if $\eta c \leq 0.68$ and $b_\eta := 1.3$ if $\eta c \leq 0.8$ (the constants for which $\log(1+x) \geq x - \beta x^2$ on $x \geq -\eta c$). Deployed, all six grid points qualify with $\beta = 1$, since $\eta_{\max} c = 0.4 \leq 0.68$. Fix any nonempty $G \subseteq \mathcal{H}$, and set $S_G := \sum_{\eta\in G}\eta$, $q := |G|$, and*

$$\eta_{\text{eff}}(G) \;:=\; \frac{1}{S_G}\sum_{\eta\in G} b_\eta\,\eta^2, \qquad \Lambda(G) \;:=\; \frac{q}{S_G}\,\log\frac{Q}{q}.$$

Then, on any sample path on which the mixture test has not yet fired,

$$\sum_{i\leq n} Z_i \;\leq\; \underbrace{\tfrac{q}{S_G}A'_k \cdot 1\{\textit{event-terminated}\} \;+\; \Lambda(G)\cdot 1\{\textit{floor-terminated}\}}_{\textit{run overhead}} \;+\; \eta_{\text{eff}}(G)\,V_n, \qquad (15)$$

and trivially $\sum_{i\leq n} Z_i \leq c\,n$. Two choices of G are worth stating. (i) *$G = \mathcal{H}$ (all grid points): $\Lambda(\mathcal{H}) = 0$ (floor restarts, however many, cost nothing, exactly as in the fixed-fraction analysis), at the cost of the grid-averaged coefficient $\eta_{\text{eff}}(\mathcal{H}) = \sum_\eta b_\eta\eta^2 / \sum_\eta \eta = 0.138$ deployed.* (ii) *$G = \{\eta\}$ (a single grid point): the tight rate $\eta_{\text{eff}} = \eta$ (at $\eta = 0.05$ and the measured variation this is ≈ 0.09 nats per probe), at the cost of $\Lambda = \log Q/\eta$ (36 nats) per floor-terminated run, because the deployed rule restarts all components together at the mixture's floor, where an individual component may still be as high as $\log Q$. Intermediate G interpolate (e.g. $G = \mathcal{H} \setminus \{0.2\}$: $\eta_{\text{eff}} = 0.070$, $\Lambda = 4.9$ nats per floor restart). We take the best of the three bounds. Only (i) and the trivial bound are free of realized counts. Which events terminate a family's runs is a property of the gate: the harm family is reset by rollbacks, breaker trips, virtual-probe reflex rollbacks, and demotions but not by commits: its null is per-bet against the anchor in force, which is $\mathcal{F}_{j-}$-measurable, so Lemma 2 covers runs that span commits, and a forced commit cannot refresh an adversary's threshold budget. The gate logs the event-terminated run count. The gain and staleness families are also reset at commits (their nulls reference the anchor).*

Proof. Write $\Lambda^\eta := \sum_{i\leq n}\log(1+\eta Z_i)$ for the run's log-wealth of component η. Since the mixture has uniform weights, $\log E^f = \log\sum_\eta e^{\Lambda^\eta} - \log Q$. Not having fired means $\log E^f < A_k$, i.e. $\sum_\eta e^{\Lambda^\eta} < Qe^{A_k}$, hence $\Lambda^\eta < A_k + \log Q = A'_k$ for every η. A floor-terminated run has $\log E^f \leq 0$, i.e. $\sum_\eta e^{\Lambda^\eta} \leq Q$, hence for any G, by Jensen applied to the q terms of G (the remaining terms are nonnegative), $\frac{1}{q}\sum_{\eta\in G}\Lambda^\eta \leq \log\left(\frac{1}{q}\sum_{\eta\in G} e^{\Lambda^\eta}\right) \leq \log(Q/q)$. Next, $\log(1+x) \geq x - \beta x^2$ for all

$x \geq -a$ whenever $\beta \geq \big(\log \frac{1}{1-a} - a\big)/a^2$: the function $g(x) = \log(1+x) - x + \beta x^2$ satisfies $g(0) = 0$ and $g'(x) = x\big(2\beta - \frac{1}{1+x}\big)$, so g increases on $[-a, \frac{1}{2\beta} - 1]$, decreases to $g(0) = 0$, and increases thereafter. Hence $g \geq 0$ on $[-a, \infty)$ as soon as $g(-a) \geq 0$, which is the displayed condition. Numerically $a = 0.68$ gives $\beta = 0.993 \leq 1$ and $a = 0.8$ gives $\beta = 1.264 \leq 1.3$. So for each $\eta \in G$, $\eta \sum_i Z_i - b_\eta \eta^2 V_n \leq \Lambda^\eta$. Averaging these q inequalities with weights proportional to η (i.e. dividing the η-th by η and averaging with weight η/S_G, which makes the left-hand sides identical) gives $\sum_i Z_i \leq \frac{1}{S_G} \sum_{\eta \in G} \Lambda^\eta + \eta_{\text{eff}}(G) V_n$, and the two bounds on $\sum_{\eta \in G} \Lambda^\eta$ above give (15). The trivial bound is $Z_i \leq c$ per probe. □

The bound is deterministic, so no detection-failure probability has to be budgeted. Its only data-dependent quantity is the observable V_n, which the gate accumulates online. The certificate reports its own tightness. Bounding $V_n \leq c^2 n$ recovers a fixed-rate (range-based) allowance. The measured $V_n/n = 1.81$ on this benchmark (against $c^2 = 4$) makes the allowance $2.2\times$ tighter, and is what keeps instantiation (i) non-trivial ($0.138 \times 1.81 \approx 0.25$ nats per probe versus the trivial $c = 2$). The same lemma applies to the staleness family (anchor vs. passive), whose grid and hence η_{eff} are the same. The gain family is analyzed directly through its log increments in Theorem 2 and needs no $x - \beta x^2$ bound. Summing (15) over runs multiplies the $\frac{q}{S_G} A'_k$ term by the number of *event-terminated* runs only.

Lemma 4 (Predictable-rate sampling)**.** *Let $s_t \in [-c, c]$ be any process adapted to $\mathcal{F}_{t-1}$, and let $B_t \sim \text{Bernoulli}(q_t)$ be coins whose rates $q_t \geq q_{\min} > 0$ are $\mathcal{F}_{t-1}$-measurable and whose outcomes are independent of $\mathcal{F}_{t-1}$ and of the environment. Write $K_{\max} := 1/q_{\min}$ and $L := \log(\pi^2 T^2/(3\delta'))$. Then with probability $\geq 1 - \delta'$, for all T simultaneously,*

$$\sum_{t \leq T} s_t \;\leq\; \sum_{t \leq T} q_t^{-1} B_t \, s_t \;+\; c\sqrt{2 K_{\max} T L} \;+\; \tfrac{2}{3}\,(K_{\max}{+}1)\, c\, L.$$

The lemma applies at every interval: the certified base-coin rate $q_t = 1/K_t$ is positive throughout, trial windows included (the base stream runs through trials, and trial draws are a second, separate stream feeding only the gain family), and it is used descriptively, never chained into a global a-priori display (App. A.12): with s_t the applied system's winsorized excess at its live intervals and B_t the harm family's bet indicator, it says the logged ledger's sample is representative of unsampled conduct up to the martingale remainders, which rules out duty-cycling (an adapter cannot confine harm to unprobed intervals). The certified rate set is $\{1/K, 1/K_{\text{slow}}\}$, each $\mathcal{F}_{t-1}$-measurable (alert counter and trust state are set by the end of the previous interval). Trial draws at rate $1/P_{\text{tr}}$ are extra, carry no certified bets, and never displace a certified coin.

Proof. $D_t = s_t - q_t^{-1} B_t s_t$ is a martingale difference w.r.t. the filtration generated by $(\mathcal{F}_{t-1}, B_{1:t})$: q_t is $\mathcal{F}_{t-1}$-measurable, so $\mathbb{E}[q_t^{-1} B_t s_t \mid \mathcal{F}_{t-1}] = s_t$. Increments obey $|D_t| \leq (K_{\max}{+}1)c$ and predictable variance $\mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}] = s_t^2(q_t^{-1} - 1) \leq c^2(K_{\max} - 1)$, and Freedman's inequality (Freedman, 1975) plus a $\delta' \cdot 3/(\pi^2 T^2)$ stakes-over-horizons union gives the time-uniform statement. □

Proposition 3 (full statement). *Group the harm family's sampled per-probe statistics into its betting runs (delimited by floor restarts and event resets, Lemma 2). Within a run let n be the sampled-bet count and $V_n := \sum_{j \leq n} \bar{Z}_j^2$ the realized quadratic variation. Pathwise and deterministically (no failure probability enters), on every prefix of a run on which the mixture has not fired,*

$$\sum_{j \leq n} \bar{Z}_j \;\leq\; \min\Big\{\, c\,n,\;\; \eta_{\text{eff}}(\mathcal{H})\, V_n + \tfrac{Q}{S_{\mathcal{H}}}\, A'_k,\;\; \min_{\eta \in \mathcal{H}} \big[\, \tfrac{A'_k}{\eta} + b_\eta\, \eta\, V_n \,\big] \Big\}$$

(the instantiations $G = \mathcal{H}$ and $G = \{\eta\}$ of Lemma 3, plus the trivial per-bet bound). Summed over the runs through time T: with N_{ev} the logged number of event-terminated runs (the harm family is reset by rollback-like events, never by commits), N_{fl} the logged floor-restart count, $k_{\max}$ the largest realized restart index (so every threshold is at most $A'_{k_{\max}}$, the crude a-priori bound $k_{\max} \leq 2T$

gives $A'_{k_{\max}} \le \log \frac{\pi^2}{2\delta} + 2\log(1+2T) + \log Q$*),* $N_{\mathrm{samp}} := \sum n$ *and* $V_T := \sum V_n$*,*

$$\sum_j \bar{Z}_j \;\le\; \Xi_T \;:=\; \min\Big\{ c\,N_{\mathrm{samp}},\;\; \eta_{\mathrm{eff}}\,V_T + \tfrac{Q}{S_{\mathcal{H}}}\,(N_{\mathrm{ev}}+1)\,A'_{k_{\max}},$$
$$\min_{\eta\in\mathcal{H}}\Big[\tfrac{(N_{\mathrm{ev}}+1)\,A'_{k_{\max}} + N_{\mathrm{fl}}\log Q}{\eta} + b_\eta\,\eta\,V_T\Big]\Big\}.$$

Every quantity on the right (N_{samp}, V_T, N_{ev}, N_{fl}, $k_{\max}$) is logged by the gate, which also accumulates the left-hand side $\sum_j \bar{Z}_j$, so both sides of the inequality are available online for each deployment (Fig. 3, and App. C reports both sides per plant, also under adversarial proposals). By Lemma 4 the sampled statistics are an unbiased martingale sample of the applied system's winsorized excess ς_t at its live intervals, at the predictable per-tier rates and at every interval, even during trial windows, since the certified base coins never pause, so the ledger covers a representative sample of conduct and a proposal cannot create an unsampled window. We do not chain it into a global a-priori display over all intervals (App. A.12).

Proof. The per-run inequality is Lemma 3: at $G = \mathcal{H}$ the run overhead is $\frac{Q}{S_{\mathcal{H}}}A'_k$ for a run that has not yet floored ($\Lambda(\mathcal{H}) = 0$, so floor-terminated runs are free at this reading). At $G = \{\eta\}$ it is A'_k/η for an unfloored run and $\log Q/\eta$ for a floor-terminated one. The trivial bound is $\bar{Z}_j \le c$ per bet. Summing over runs, each event-terminated run and the run in progress contribute one threshold overhead each, each floor-terminated run its $\Lambda(G)$, and the V terms add to V_T. Bounding every realized A'_k by $A'_{k_{\max}}$ gives the display. All named quantities are maintained online by the gate implementation included in the supplementary material. □

Remark (readings, deployed). At the deployed grid the allowance rate is $\eta_{\mathrm{eff}}\bar{v} = 0.138 \times 1.81 \approx 0.25$ nats per sampled probe against the trivial $c = 2$. The single-fraction reading at $\eta = 0.05$ tightens the rate to $0.05\,\bar{v} \approx 0.091$ at $\log Q/\eta = 36$ nats per floor restart. When floor restarts and rollback-like events are both rare (both logged), the grid-optimized branch reads $2\sqrt{(N_{\mathrm{ev}}+1)\,A'_{k_{\max}}V_T} = \tilde{O}(\sqrt{T})$, the Proposition A(b) rate down to the grid floor of Proposition A(a5). The ledger is this tight because the measured realized variation (App. B.4) is $2.2\times$ below the range bound c^2.

A.9 Assumptions (restated)

Assumption 1′ (bounded amplification, high probability). The event of Lemma 1 holds (probability $\ge 1 - \delta_n$, $\delta_n = Te^{-n_I/16}$).

Assumption 2 (paired conditional drift, winsorized form). For adjacent intervals, every fixed filter θ, and every $\mathcal{F}_{t-1}$-measurable block vector $A \in \mathbb{R}^m$,

$$\Big|\mathbb{E}\Big[m^{-1}\textstyle\sum_b \big(\mathrm{clip}(A_b - \log \ell_t^{(b)}(\theta), \pm c) - \mathrm{clip}(A_b - \log \ell_{t-1}^{(b)}(\theta), \pm c)\big)\;\Big|\;\mathcal{F}_{t-1}\Big]\Big| \;\le\; \kappa,$$

with $\kappa < \gamma/2 = 0.25$. We adopt this winsorized form as the primitive because it is the form the paired bet statistics take (they are clipped). Since clip is 1-Lipschitz, it is implied by the plain mean-drift form $|\mathbb{E}[\log \ell_{t+1}^{(b)}(\theta) - \log \ell_t^{(b)}(\theta) \mid \mathcal{F}_{t-1}]| \le \kappa$ whenever clipping is inactive, and in general by a κ-bound on the Wasserstein-1 drift of the conditional block-loss law. The assumption bounds the conditional mean (predictable drift) and says nothing about realized fluctuations. On the benchmark the realized adjacent-interval increment has standard deviation 0.445 nats (median $|\Delta| \approx 0.24$). That scatter is the martingale part of the loss process, which the betting penalty (h^*) and Lemma 4 account for, and it is consistent with any κ, however large the realized $|\Delta|$. We measure the drift per suite and per filter state, in the winsorized form above, as 32-interval windowed means of the one-step winsorized paired increment of fixed filters (passive, a converged control filter, the same filter stale after the first re-seat), replayed counterfactually on all four suites, including the polarity suite, where the frozen filter is mis-adapted, and the mid-episode fault suite, where the true S has flipped underneath it (`drift_check2.py`, App. B.3). Across all twelve suite×state cells the window medians are 0.013–0.019 nats, the 95th percentiles 0.056–0.070, and the global maximum over every measured window is 0.176 (severe-re-seat suite, passive state). Three levels are characterized (table in App. A.10). No measured window in any cell exceeds the margin $\gamma/2 = 0.25$, not even windows overlapping a crossfade or lying in a fault regime, so the margin holds uniformly

at the window level. The intermediate level 0.13 (between the worst-cell p95 and the maximum) is exceeded by at most 0.7% of windows in the worst cell (zero in most), a fraction budgeted through Assumption $2'$. Finally, $\kappa_{\max} := 0.176$ is the uniform fallback for any constant that requires a bound with no exceptions. The conjectured fault-regime blowup ($\kappa_{\text{gap}} \sim 3.4$ under faults) does not occur: a stale filter's harm *level* is large, but the temporal drift of its paired statistic is not: stale-state p95 0.067 (polarity) and 0.059 (mid-episode fault).

Assumption $2'$ (drift failures). Assumption 2 may fail on a set of at most νT intervals (regime crossfades and event onsets, where transient conditional drift concentrates). Each such interval can bias a winsorized sampled statistic by at most c. Every result that consumes a drift constant therefore quotes it together with the measured exceedance fraction ν at the consumed level (App. A.10). Measured per suite: the exceedance fraction is zero at $\gamma/2 = 0.25$ and at 0.21 in every cell, and at most 0.7% of windows at the 0.13 level (worst cell, zero in most), so $\nu c \leq 0.014$ nats per interval at that level. Crossfade intervals ($< 2\%$ of the total, App. B) upper-bound the term at $\nu c < 0.04$, and we retain the larger figure as a safeguard.

A.10 Measured drift levels and the constants that consume them

The per-suite winsorized drift measurement behind Assumption 2 (App. B.3) yields three levels: the worst-cell window median 0.019 (typical), the worst-cell window 95th percentile 0.070 (p95, the $\leq 0.7\%$ of windows above it are budgeted by Assumption $2'$), and the global window maximum 0.176. With the a-priori route set aside (App. A.12), exactly two results consume a drift constant: Theorems 1 and 2, both through the in-trial pairing gap $\kappa_{\text{tr}} = \kappa\,\bar{\Gamma}^{\text{tr}}$, instantiated at the high-probability gap constant $\bar{\Gamma}^{\text{tr}} = 2/q_{\text{tr}} \approx 13.3$ intervals that their proofs derive (the mean gap, ≈ 6.7, is quoted descriptively only):

constant (margin)	role	at p95 drift	at median drift
κ, one step ($< \gamma/2 = 0.25$)	harm pairing bias (interpretive)	no measured window exceeds the margin, in any cell	
κ_{tr} ($< \gamma = 0.5$)	consumed: Thms. 1, 2	≈ 0.93: exceeds γ, no margin	≈ 0.26 (margin 0.24)
$\gamma_s = 0.25$ vs. reference-gap drift	demotion margin (descriptive)	≈ 0.75: exceeded	≈ 0.20: clears
fault-regime drift	conjectured ~ 3.4 blow-up	stale-state p95 ≤ 0.067: refuted	

Both consumers instantiate $\bar{\Gamma}^{\text{tr}}$ at the high-probability constant $2/q_{\text{tr}}$ their proofs derive. At that constant the certified margin is positive under median-level measured drift (0.24 nats) and not at the p95 level, where, as under drift sustained at the global maximum, a commit certifies no margin and protection reverts to the harm ledger, the validation certificate and the reflexes. The one-step κ enters no conclusion (the ledger is pathwise on the sampled statistic) but calibrates how each sampled $\bar{Z}_j$ relates to the pre-probe paired gap (one application of Assumption 2). The staleness margin γ_s is *descriptive by design*: no demotion-counting theorem is claimed. The margin is sized so that typical measured reference-gap drift at the full probe rate (≈ 0.20 nats per bet at the median level) cannot fire the staleness test alone, while gap-drift excursions above it, and the $4\times$ longer reference gaps of the thinned tier, can fire only *extra* demotions, which move the system toward the invariant fallback: the safe direction. The conjectured fault-regime blow-up is refuted by measurement: a stale filter's harm *level* is large, but the temporal drift of its paired statistic is not.

A.11 The per-trial commit certificate

Theorem 1 (full statement). *Fix any adaptation process π, adversarial and aware of EGA's internals but not of the gate's future coins, and any calibration model $\hat{S}$. Define the in-trial pairing-drift constant*

$$\kappa_{\text{tr}} \;:=\; \kappa\,\bar{\Gamma}^{\text{tr}}, \qquad \bar{\Gamma}^{\text{tr}} \;\leq\; \tfrac{2}{q_{\text{tr}}} \;\; \textit{h.p.}, \qquad q_{\text{tr}} := \tfrac{1-1/J}{P_{\text{tr}}} = 0.15,$$

where $\bar{\Gamma}^{\text{tr}}$ is the per-bet average of the realized gaps Γ_j between a trial bet's probe and its stored opposite-side measurement: successive non-zero trial probes arrive with predictable per-interval probability q_{tr}, so each Γ_j is dominated by $\text{Geometric}(q_{\text{tr}})$, and the proof's concentration step controls $\bar{\Gamma}^{\text{tr}}$ only at the high-probability constant $2/q_{\text{tr}} \approx 13.3$ intervals (the mean gap $1/q_{\text{tr}} \approx 6.7$ is reported for context only, and the gaps are coin-driven, independent of the environment). Every numerical instantiation of κ_{tr} below therefore uses $\bar{\Gamma}^{\text{tr}} = 2/q_{\text{tr}}$. Then:

(a) (Each trial is a fresh e-process on a fixed pair.) *A commit trial runs its own restarted gain wealth (reset to* 1 *at the proposal, with its own restart index in the spending schedule) whose every bet pairs the* same two gate-stored objects*: the frozen snapshot copy* W_{tr} *and the anchor* a*, neither of which the adaptation process can influence once the trial has started (a rollback-like event kills the trial, and the live candidate's updates touch neither object). A threshold crossing therefore certifies a property of* exactly the object a commit promotes*, at the trial's own* δ_k*: pooling gain wealth across a* changing *sequence of snapshots would certify none of them. The frozen-pair design closes that gap by construction.*

(b) (Both-direction bets, one bet per realization, validity.) *At an anchor-side trial probe the bet statistic* $\bar{Z}_j^{\mathrm{tr}}$ *is the block-mean of* $\mathrm{clip}(\log \ell^{(b)}(W_{\mathrm{tr}}, \textit{stored}) - \log \ell^{(b)}(a, \textit{now}), \pm c)$*. At a snapshot-side probe the roles reverse (*W_{tr} *measured now, anchor stored). Every realization enters exactly one bet, stored or fresh: a probe's measurement either pairs with the stored opposite-side measurement, in which case that single bet consumes both (the implementation invalidates the store on use, and a measurement spent in a bet is never itself stored), or it is stored for the next opposite-side probe, never both. Bets therefore fire at every other non-zero trial probe (*$\approx n/2$ *per* n*), no realization is ever re-bet, and neither consecutive-probe multiplicity nor double-counting (one measurement entering two sign-aligned bets, as the fresh side of its own bet and the stored side of the next) can arise. The stored side is* $\mathcal{F}_{j-}$*-measurable at bet time and the trial schedule and consumption flags are* $\mathcal{F}_{t-1}$*-measurable, so Lemma 2 applies as stated to the per-bet conditional null.*

(c) (What a commit certifies: pairwise and per event.) *On the intersection of the validity and concentration events of the proof (jointly* $\geq 1 - \delta/3 - \delta_g$ *over all trials), every commit attests that the fixed pair satisfied, in the run-average conditional-mean sense, a winsorized improvement of* W_{tr} *over the outgoing anchor of at least* $\gamma - \kappa_{\mathrm{tr}}$ *nats at commit time, for* any π *and any* $\hat{S}$*. The statement is* pairwise and per event*: it compares the promoted object to the outgoing anchor at that trial's* δ_k*, and it is chained into no cross-commit accounting: no ratchet count, no anchor-vs-passive transfer (App. A.12).*

Proof. (a) is the trial lifecycle: W_{tr} is a copy frozen at the proposal (an $\mathcal{F}$-measurable functional of past data and $\hat{S}$). The gain family is reset there with an incremented restart index, so the trial's wealth is a restarted e-process in the sense of Lemma 2, drawing on the gain family's $\delta/3$ spending schedule. The anchor can change during a trial only through a commit (which ends it) or a demotion (a rollback-like event, which kills it), so the pair is fixed over the trial's bets.

(b)–(c): Under the null (the fixed pair's same-time winsorized conditional gap is $\geq -\gamma + \kappa_{\mathrm{tr}}$ throughout the trial), write each bet's conditional mean as $\mu_j + \Delta_j$: $\mu_j \leq 0$ collects the same-time gap plus the predictable drift over the realized gap (at most $\Gamma_j \kappa$, Assumption 2 telescoped over the gap), and Δ_j is the stored realization's deviation from its own conditional mean, a martingale difference across bets, because one-bet-per-realization makes every fluctuation, stored or fresh, enter exactly one bet. Two concentration steps remain. *Gap total:* the Γ_j are dominated by independent $\mathrm{Geometric}(q_{\mathrm{tr}})$ variables (trial coins are data-independent at a predictable rate). For such a variable $\mathbb{E}[e^{sG}] \leq 2$ at $e^s = 1/(1 - q_{\mathrm{tr}}/2)$, so $\mathbb{P}\big(\sum_{j \leq n} \Gamma_j > \frac{2}{q_{\mathrm{tr}}}(n+t)\big) \leq 2^n e^{-(n+t)} \leq e^{-t}$. With $t = \log(2/\delta_g)$ this bounds the per-trial drift charge by $\kappa \cdot (2/q_{\mathrm{tr}})(n + \log(2/\delta_g))$, i.e. κ_{tr} per bet up to the stated h.p. factor. *Fluctuation total:* the Δ_j form a martingale with increments bounded by $2c$ over the at most n_{tr} bets of any trial. Azuma–Hoeffding per trial, with a union over trials at total budget $\delta_g/2$ (spending proportional to $1/(k+1)^2$ over trial restart indices k), bounds it uniformly. *Union over trials (wealth crossings):* the trial started at restart index k of the gain family bets against its own threshold $1/\delta_k$ with $\delta_k = \frac{\delta}{3} \cdot \frac{6}{\pi^2 (k+1)^2}$, and every trial strictly increments the index. Ville's inequality per trial and a union over all trials (however many occur, the count may be unbounded) costs at most $\sum_{k \geq 0} \delta_k = \delta/3$ in total, so the spending schedule already carries the union. The wealth against μ_j alone is a nonnegative supermartingale (Lemma 2). A crossing under the null therefore has probability at most $\delta/3 + \delta_g$ over all trials, and on the complement every crossing attests a run-average conditional gap $< -(\gamma - \kappa_{\mathrm{tr}})$, which is (c). □

Remark (measured instantiation of κ_{tr}). At the high-probability gap constant the proof derives ($\bar{\Gamma}^{\mathrm{tr}} = 2/q_{\mathrm{tr}} \approx 13.3$ intervals), the per-suite drift measurement (App. A.10) gives $\kappa_{\mathrm{tr}} \approx 0.26$ at the worst-cell window median (typical drift) for a certified margin of $\gamma - \kappa_{\mathrm{tr}} \approx 0.24$ nats ($\approx$1.1 dB), while at the worst-cell window p95 level $\kappa_{\mathrm{tr}} \approx 0.93$ exceeds $\gamma = 0.5$ and a commit certifies no

margin: the certificate's quantitative margin holds under median-level measured drift, not at p95, and the measured drift table of App. A.10 remains the reference. Two roles must not be conflated: coin *unpredictability* makes the logged ledger's samples representative of unsampled conduct (Lemma 4). Commit validity needs only *predictability* of the bet times, so model-*proposed* trials accelerate commits without weakening what a commit certifies: a wrong model can only waste trials, never promote an uncertified object.

A.12 What is not proved: the a-priori route

A fully a-priori cumulative-harm display would chain the per-event certificates: a ratchet bound on commit counts, cooloff telescoping for rollback counts, demotion counting, occupancy budgets and reference-gap transfer for an anchor-vs-passive second tier, and a tiered sampling-transfer ledger over all intervals. Each link needs its own assumption (piecewise stationarity of the environment, uniform drift transfer across multi-interval reference gaps, occupancy bounds), and with the measured constants the assembled bound exceeds the trivial winsorization cap cT at every deployed horizon, the streak-decay backoff alone contributing an event term linear in T. We do not state it. The certified results are validity of the betting families (Lemma 2), the logged harm ledger, audited online (Proposition 3), the per-trial commit certificate (Theorem 1), their independence of the model (Proposition 2), the validation certificate (Proposition 1), and adoption (Theorem 2). A non-vacuous a-priori cumulative-harm account at deployment horizons (or a proof that none exists for supervisors of this class) is an open problem. The drift measurements of App. B.3 remain as empirical characterization of the pairing margins that the results above consume (App. A.10).

A.13 Full statement and proof of Theorem 2

Theorem 2 (full statement). *Suppose from some interval onward the candidate maintains a conditional winsorized improvement over the anchor of at least $\beta \in (\gamma, c]$ at its live intervals: the margin must clear the commit margin $\gamma = 0.5$ (positive drift requires $\beta > \gamma$, plus the second-order slack below), and $\beta \le c$ by winsorization. Assume also that neither the harm test (valid w.p. $1-\delta$, since the candidate is genuinely better) nor the breaker fires during the adoption window (the breaker cannot: a candidate satisfying the premise is below the anchor's level, while the breaker only trips $c_{\mathrm{brk}} > 0$ nats above it). Since commits are exclusively trial-certified, adoption proceeds in two phases.*

Phase 1 (proposal, model-assisted, conditional on calibration). *The virtual gain wealth bets on the virtual statistic at every eligible live interval. If the virtual channel is b_v-calibrated (conditional bias at most b_v, which Proposition 1 tests online) and $\beta - b_v > \gamma$ plus the slack below, the drift calculation of Phase 2 applied with margin $\beta - b_v$ opens a trial within $\approx A'_k/\varepsilon_0$ live intervals, interval-rate evidence, independent of the probe schedule K_t. Adoption therefore depends on the model in a way safety does not, since a distrusted or miscalibrated model may never trigger a proposal, and adoption then stalls.*

Phase 2 (trial, unconditional given the premise). *A trial's non-zero probes arrive at the predictable rate $q_{\mathrm{tr}} = (1 - 1/J)/P_{\mathrm{tr}} = 0.15$ per interval and its bets at $\approx q_{\mathrm{tr}}/2 = 0.075$ (one bet per two non-zero trial probes under one-bet-per-realization, $\approx 2/(1-1/J) \approx 2.7$ trial draws, i.e. $\approx$13.3 intervals, per bet, zero probes interleave without advancing the alternation), on the fixed pair (W_{tr}, a), with bet-level margin $\beta_{\mathrm{tr}} := \beta - \kappa_{\mathrm{tr}}$ (the in-trial pairing drift of Theorem 1 at its high-probability gap constant). The latency figures below use the median-drift instantiation $\beta_{\mathrm{tr}} \ge \beta - 0.26$. At p95-level drift $\kappa_{\mathrm{tr}} \approx 0.93$ leaves no certified margin and marginal candidates fail sooner, consistent with the failed single-trial hypothesis reported below. The trial wealth is the mixture* (6) *over $\mathcal{H}_{\mathrm{gain}}$ ($Q_{\mathrm{gain}} = 5$) on $Z'_j = -\bar{Z}^{\mathrm{tr}}_j - \gamma \in [-(c+\gamma), c-\gamma]$. All wealth factors are positive since $\max \mathcal{H}_{\mathrm{gain}}\,(c+\gamma) = 0.875 < 1$. It suffices to track any single grid point, whose wealth the mixture dominates up to $1/Q_{\mathrm{gain}}$. The threshold to clear is $A'_k := A_k + \log Q_{\mathrm{gain}}$. Define the per-point drift constant via the actual log increment (this is the definition of ε_0 the main text refers to, quoting the value at the best valid grid point), with β_{tr} in place of β:*

$$\varepsilon_0(\eta) \;=\; \log\bigl(1 + \eta(\beta_{\mathrm{tr}} - \gamma)\bigr) \;-\; \frac{\eta^2\, v}{2\bigl(1 - \eta(c+\gamma)\bigr)^2}, \tag{16}$$

where v bounds the conditional variance of $\bar{Z}^{\mathrm{tr}}_j$: the first term is the increment at the premise level and the second is the (second-order) Jensen correction for the block-mean's fluctuations. No $x - x^2$

bound is needed for the gain family. Call η valid *if* $\eta(c+\gamma) < 1$ *and* $\varepsilon_0(\eta) > 0$*. Increments are bounded by*

$$M_0(\eta) \;=\; \max\Big(\big|\log\big(1-\eta(c+\gamma)\big)\big|,\; \log\big(1+\eta(c-\gamma)\big)\Big).$$

Then for any valid η *and any* $\delta' \in (0,1)$*, a trial started at restart* k *crosses within*

$$N \;=\; \frac{2A'_k}{\varepsilon_0(\eta)} + \frac{8M_0(\eta)^2}{\varepsilon_0(\eta)^2}\log\frac{1}{\delta'} \quad \textit{trial bets, w.p. } \geq 1-\delta-\delta',$$

where $A_k = \log(\pi^2(k+1)^2/(2\delta))$ *and* k *is the (data-dependent) restart index at trial start. The bound may be optimized over valid grid points. Because a failed trial resets the gain wealth, a single trial certifies the commit only if* $N \leq n_{\mathrm{tr}}$*, the trial's bet budget. This is a hypothesis of the deployed instantiation and is checked below.*

Deployed hypothesis check (each constant against its own hypothesis, at $\beta_{\mathrm{tr}} = c$*,* $v = \bar{v} = 1.81$***).*** *The Jensen denominator* $1-\eta(c+\gamma)$ *and the resulting* $\varepsilon_0(\eta)$ *across the gain grid:* $\eta = 0.02$*:* 0.029*,* 0.05*:* 0.069*,* 0.1*:* $\log 1.15 - 0.0181/1.125 = 0.1398 - 0.0161 = 0.124$*,* 0.2*:* $0.262 - 0.145 = 0.118$*,* 0.35*:* $0.422 - 7.10 < 0$*, so the largest fraction fails its own hypothesis (its Jensen denominator is* 0.125*, squared* 0.0156*) and the theorem cannot be instantiated there. This costs the mixture only its* $\log Q_{\mathrm{gain}}$ *nats, and the first-order term at* $\eta = 0.35$ *must not be quoted on its own. The latency bound is best at* $\eta = 0.1$ *(*$\varepsilon_0 = 0.124$*,* $M_0 = |\log 0.75| = 0.288$*,* $M_0^2/\varepsilon_0^2 = 5.4$*), giving* $N \approx 100 + 43\log(1/\delta')$ *trial bets. At the deployed trial length this hypothesis fails:* $n_{\mathrm{tr}} = 24$ *bets supply in-expectation wealth* $24 \times 0.124 \approx 3.0$ *nats against* $A'_0 = \log(\pi^2/(2\delta)) + \log 5 \approx 6.2$*, so the high-probability single-trial guarantee does not hold for a marginal candidate at the measured variance: it would require* $n_{\mathrm{tr}} \approx 100$*, or conditional variance* $v \lesssim 0.04$*. At the deployed constants the certified statement is weaker: a strongly better candidate (*$\bar{Z}_j^{\mathrm{tr}}$ *pinned near* $-c$*, so* $Z'_j \approx c - \gamma = 1.5$ *with negligible conditional variance) drives the* $\eta = 0.35$ *component at* $\log(1 + 0.35 \times 1.5) = 0.42$ *nats per bet and crosses in* $\lceil A'_0/0.42 \rceil = 15 \leq n_{\mathrm{tr}}$ *bets. Marginal candidates rely on repeated trials (each failure costs the* $P_{\mathrm{cool}} = 192$*-interval cooldown,* ≈ 3.1 *s), whose per-trial success probability this argument does not bound from below.*

Wall-clock reading. *For a clearly better candidate: proposal within* $\approx A'_0/\varepsilon_0 \approx 50$ *live intervals (*≈ 0.8 *s at 16 ms intervals, interval-rate virtual evidence, requires model calibration) plus a trial crossing in* ≈ 15 *bets at the bet rate* $q_{\mathrm{tr}}/2 \approx 0.075$*/interval (*$\approx 200$ *intervals,* ≈ 3.2 *s):* ≈ 4.0 *s end-to-end, independent of the thinned schedule* K_t *(the trial rate is* $1/P_{\mathrm{tr}}$ *in either tier). The ratchet then retains the improvement until the next plant change. Improvements with margin below* γ *are not adopted (by design), and the* $\log Q_{\mathrm{gain}}$ *nats the mixture costs here are repaid on the harm side, where the grid shrinks the fixed-fraction sensitivity floor* $40\times$ *(Proposition A).*

Proof. Fix a valid η. The increments $X_j = \log(1+\eta Z'_j)$ are well defined and lie in $[\log(1-\eta(c+\gamma)), \log(1+\eta(c-\gamma))] \subset [-M_0, M_0]$. For their conditional mean, second-order Taylor of $f(x) = \log(1+\eta x)$ around the conditional mean of Z'_j gives $\mathbb{E}[f(Z'_j) \mid \mathcal{F}] \geq f(\mathbb{E}[Z'_j \mid \mathcal{F}]) - \frac{1}{2}\sup|f''| \cdot v$ with $\sup|f''| = \eta^2/(1-\eta(c+\gamma))^2$ on the statistic's range. The trial pair is fixed, so the premise transfers to both bet directions: by the decomposition of Theorem 1(b), each bet's conditional mean satisfies $\mathbb{E}[Z'_j \mid \mathcal{F}] \geq \beta - \kappa_{\mathrm{tr}} - \gamma = \beta_{\mathrm{tr}} - \gamma$ up to the stored-realization martingale part, which the Azuma step below absorbs. Since f is increasing, $\mathbb{E}[X_j \mid \mathcal{F}] \geq \varepsilon_0(\eta)$. The premise concerns the conditional mean given everything before the bet, so predictable trial bet times are covered (Lemma 2). The centered partial sums form a martingale with increments bounded by $2M_0$. Azuma–Hoeffding over the first N bets with $N\varepsilon_0 \geq 2A'_k +$ slack gives the tail (the bound is applied at the single deterministic time N, not uniformly, so no union is needed). Since the mixture wealth is at least $1/Q_{\mathrm{gain}}$ times this component's, crossing A'_k for the component implies crossing A_k for the mixture. The argument uses both parts of the validity of η: $\eta(c+\gamma) < 1$ keeps X_j finite and $\varepsilon_0(\eta) > 0$ makes N finite. Both are checked numerically above at the deployed constants, as is the single-trial condition $N \leq n_{\mathrm{tr}}$ (which fails for marginal candidates and holds for strongly better ones, reported above). Phase 1 is the same calculation applied to the virtual statistic with margin $\beta - b_v$, and its bets ride every eligible live interval, hence the interval-rate latency. The ratchet retains the improvement until the next plant change. □

A.14 Remarks

(i) Randomization. With deterministic probes an adapter aware of the schedule can confine harm to unprobed intervals ("duty-cycling"), making sampled harm zero while true harm grows linearly. Lemma 4 says Bernoulli coins prevent this: the sampled statistics are an unbiased martingale sample of the process's winsorized excess, uniformly over adversaries and horizons. A predictable rate $1/K_t$ does not weaken this, since the adapter may know the rate but not the coin outcome. The rate only rescales the per-tier sampling rates reported in the logs. *(ii) The object the commit certificate attaches to.* The candidate evolves, so the certificate must be attached to a specific filter. At a proposal the gate freezes a copy W_{tr} of W_{snap} (the candidate as of the last non-probe interval), collects every gain bet of the trial on the fixed pair (W_{tr}, a) (the live candidate keeps adapting, at no cost to the certificate), and a commit installs W_{tr}, not the live candidate. The measured object and the promoted object coincide identically, and the measured-sequence-vs-committed-object gap is closed *by construction*: no commit-transfer assumption enters, and Theorem 1 pays only the in-trial pairing drift κ_{tr}. *(iii) Hot-swap transients.* The first $\lceil L_S/H_b \rceil$ blocks of every probe interval (the full ring-out of the pre-swap drive) are discarded from all statistics. The residual leakage is zero by linearity of the plant and finite support of S. *(iv) Virtual probes.* The virtual probe $\hat{e}(a) = e - \hat{S} * \sigma(W{*}x) + \hat{S} * \sigma(a{*}x)$ is computable on hardware from the measured error and the commanded filters. Its error is $(S - \hat{S}) * (\sigma(W{*}x) - \sigma(a{*}x))$, so it scales with the filter difference rather than the signal level. Measured on the benchmark, the paired virtual statistic has bias median 0.007 / p95 0.071 nats before a re-seat, but p95 0.81 after re-seats and median -1.33 under the polarity fault. A pre-registered test of virtual probes as a blanket replacement for real probes failed (pooled p95 0.64), which is why they are confined to the trust-gated roles above (proposal, reflex, and validation), and why no virtual quantity appears in any certified wealth (Proposition 2). The certificate's evidence is, and remains, real paired measurements at coin-scheduled times.

B Environment Details

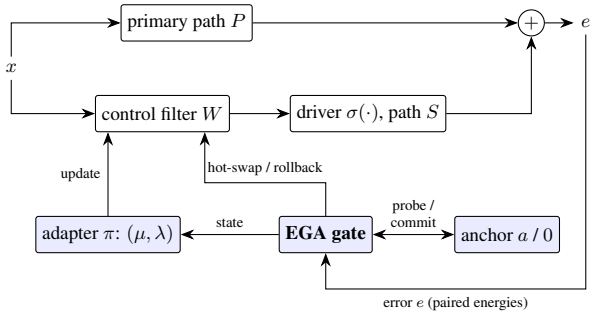


Figure 2: Feedforward ANC plant (top) and the EGA supervisor (bottom): the adapter drives the FxNLMS update. The gate probes the anchor at random times and accumulates paired evidence that drives its rollback, commit, and demotion decisions.

B.1 Acoustic plant generation

All signals run at $f_s = 16\,\mathrm{kHz}$. Adaptation blocks are $H_b = 32$ samples (2 ms). Control intervals are $M = 8$ blocks (16 ms), and the control filter has $L = 128$ taps.

Primary paths P (external noise at the reference microphone $\rightarrow$ error microphone, 256 taps): a delayed, exponentially decaying noise tail (decay 1–3 ms) with a dominant direct component, band-limited to $[30, 1500\text{–}4000]$ Hz (passive isolation), shaped by 2–4 random biquad resonances ($Q \in [0.8, 3]$, gain ± 4 dB), delay 24–56 samples (1.5–3.5 ms), and overall gain -6 to 0 dB. Broadband feedforward cancellation is causal because of this delay margin over the secondary path.

Secondary paths S (driver $\rightarrow$ error microphone, 128 taps): cepstral minimum-phase impulse responses synthesized from a smooth random log-magnitude (12 dB/oct band edges at $[30\text{–}60, 5000\text{–}7500]$ Hz plus 1–3 Gaussian resonance bumps of ± 3 dB), delayed by 4–12 samples (0.25–0.75 ms

of electro-acoustic latency), plus a -25 dB mixed-phase perturbation. Minimum-phase-plus-delay structure matches close-range driver–microphone physics and keeps the plant invertible. With it the median causal-Wiener ceiling over plants is ≈ 17 dB (Fig. 6).

Calibration. $\hat{S} = S+$ white estimation noise at 30 dB SNR (10 dB in the OOD-calibration suite), fixed for the whole episode.

Re-seating events. At 1–2 random times (steps 15%–85% of the episode), (P, S) are replaced by correlated perturbations: convex mixtures with fresh path draws (severity $s \in [0.2, 0.7]$ replaces a corresponding energy fraction), spectral delay shifts (up to ± 8 samples for P, ± 1.5 samples for S, since cavity leakage changes far more than the rigid driver–microphone geometry), and gain changes up to $\pm 4s$ dB, applied under a 300 ms crossfade. $\hat{S}$ remains stale. The OOD-severity suite uses 2–4 events with $s \in [0.7, 1.0]$.

Noise. Episode-length reference signals are regime-switching mixtures with 250 ms crossfades, segments of 2–5 s, and per-segment levels of ± 6 dB: *engine* (8-harmonic stack on a random-walk $f_0 \in [60, 300]$ Hz over a $1/f^{1.5}$ floor), *street* ($1/f$ noise with Gaussian vehicle bursts), *hvac* ($1/f^2$), and *babble* (speech-shaped, 3–6 Hz syllabic AM, held out of training and used only in the OOD-noise suite). Sensor noise: -40 dB relative to unit-RMS reference. Driver saturation: $y_{\max} = 2$. Episodes: 6 s (training), 12 s (test).

B.2 Verification of the simulator

The vectorized GPU implementation is cross-checked against a scalar, per-sample textbook FxNLMS reference on identical paths and signals (agreement within 0.1 dB), and converges to the exact causal-Wiener optimum (computed by least squares) on white noise with exact $\hat{S} = S$.

B.3 Measured drift (Assumption 2): per suite, winsorized, non-passive filters included

On held-out episodes we record paired statistics of *fixed* filters replayed counterfactually (override passes on identical episodes). The *realized* adjacent-interval increment has standard deviation 0.445 nats (median $|\Delta_t| \approx 0.24$, 95th percentile ≈ 0.92). These increments are martingale fluctuation of the loss process and do not represent drift. Both the betting penalty and the sampling term account for them (App. A). The *conditional-mean* drift is estimated with 32-interval windowed means. A raw log-loss increment measured on the main suite alone would not be the object the proofs consume. The measurement script included in the supplementary material (`drift_check2.py`) therefore measures the drift of the statistic the gate actually bets on: the winsorized ($c = 2$), post-ring-out block-mean paired increment, *per suite* (main, severe re-seat, polarity, mid-episode fault, 96 plants each, disjoint seeds) and *per filter state*. The three states are passive ($\theta = 0$), a converged control filter W_{snap} on windows before the first re-seat, and the same filter *stale* on windows after it. Under the polarity suite W_{snap} is mis-adapted, and under the mid-episode fault the true S has flipped, so these are exactly the fault-regime cells where a drift blow-up would be most plausible. Across all twelve suite×state cells, window medians run 0.013–0.019 nats, 95th percentiles run 0.056–0.070, and the global maximum is 0.176 (severe-re-seat suite, passive state). Crossfade-overlapping windows are reported both included and excluded (the difference is small, e.g. severe-re-seat stale p95 0.056 included vs. 0.046 excluded). *No window in any cell exceeds the* $\gamma/2 = 0.25$ *margin of Assumption 2, nor even* 0.21. At most 0.7% of windows in the worst cell exceed the 0.13 level, and Assumption $2'$ budgets them. Transient drift concentrates in the regime crossfades (250–300 ms, $< 2\%$ of intervals), and Assumption $2'$ budgets them too. App. A.10 lists the constants the proofs use next to these measured levels.

B.4 Measured realized quadratic variation

The harm family's realized variation per sampled probe, V_n/n, is accumulated online by the gate and logged per plant (`measure_variation.py`). Per-plant means (95th percentiles) against the range bound $c^2 = 4$: main suite, EGA-GRU 1.81 (3.58), polarity fault 1.22 (2.12), fault_mid 1.56 (3.08), and, for an *adversarial* proposal (FxNLMS $\mu = 0.2$ under the gate), 1.65 (3.41). The value is stable across benign, faulted and adversarial conditions and sits $2.2\times$ below the range, which is what

keeps the variance-adaptive allowance of the ledger (Prop. 3) non-trivial. Because V_n is observable and logged, the certificate reports this tightness at run time instead of assuming it.

B.5 GATE CONSTANTS

$\delta = 0.05$. Each of the three certified-tier e-test families (harm, gain, staleness) spends $\delta/3$ of a shared restart-indexed schedule $\delta_k \propto \delta/(3(k+1)^2)$, jointly $\leq \delta$ (Lem. 2). The validation family (App. A.7) spends an *independent*, two-sided $\delta_v = 2\delta/3$, accounted separately because its null concerns the evidence channel and not any particular filter (Sec. 5). δ_g, the auxiliary concentration parameter in the commit certificate's in-trial gap bound (App. A.11), is a separate free parameter and does not draw on δ. Betting is a *uniform mixture* over the fraction grids $\mathcal{H}_{\mathrm{harm}} = \mathcal{H}_{\mathrm{stale}} = \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$ and $\mathcal{H}_{\mathrm{gain}} = \{0.02, 0.05, 0.1, 0.2, 0.35\}$ (the fixed-fraction fallbacks used in the ablation of App. C.5 are $\eta_{\mathrm{harm}} = 0.05$, $\eta_{\mathrm{stale}} = \eta_{\mathrm{gain}} = 0.3$). The staleness margin is $\gamma_s = 0.25$ nats. This exceeds the measured full-rate reference-gap drift of $\approx$0.20 at typical levels (App. A.10), so typical measured drift alone cannot fire demotion. Winsorization is $c = 2$. The commit margin is $\gamma = 0.5$ (≈ 2.2 dB, drift margin $\gamma/2 = 0.25$). Scheduled probes are Bernoulli$(1/K_t)$ with $K_t \in \{8, 32\}$, $\mathcal{F}$-predictable via the trust machinery (trust after 6 validated probes, validation margin $\gamma_v = 0.25$, virtual-reflex margin 1 nat with patience 2). Commit trials open on virtual gain crossings: trial probe rate $1/P_{\mathrm{tr}} = 1/5$, budget $n_{\mathrm{tr}} = 24$ bets, proposal cooldown $P_{\mathrm{cool}} = 192$ intervals. The post-rollback alert window is $a_{\mathrm{al}} = 12$ scheduled probes at the fast rate (trust voided only by validation crossings and demotions). Zero probes are Bernoulli$(1/J)$ among probes with $J = 4$ (a second gate coin, so zero-probe times are unpredictable). The base cooloff is $C = 4$ full-rate probe cycles (decremented $1/K$ per interval). Doubling backoff is capped at 2^{20} cycles. No experiment approaches the cap, so the analysis treats backoff as uncapped. The backoff level is not reset but decays by one step per certified commit and by one step per $D = 32$ consecutive rollback-free scheduled probes (streak decay). Safe-side refuge sends rollbacks to $W = 0$ instead of the anchor whenever the *lagged* (pre-probe) staleness wealth exceeds 1 nat (demotion of the anchor itself still requires the full staleness certificate). The breaker margin is $c_{\mathrm{brk}} = 1.5$ nats with patience $p = 2$. The transient skip is $\lceil L_S/H_b \rceil = 4$ blocks (the full secondary-path ring-out). That leaves $m = 4$ paired blocks per probe, and the gate bets once per probe on their winsorized mean.

B.6 POLICY TRAINING

PPO (Schulman et al., 2017) with GAE(λ=0.95) (Schulman et al., 2016), $\gamma = 0.99$, clip 0.2, 4 epochs, minibatch 65,536, Adam 3×10^{-4}, entropy 10^{-3}, value coefficient 0.5, gradient clip 0.5. Actor-critic MLP $25 \to 128 \to 128$ (tanh) with Gaussian actions, $\log\sigma$ initialized to -1.2, and the action mean initialized at a conservative prior ($\mu \approx 5\times10^{-3}$, low leakage). Training uses 1024 parallel plants, 375-step on-policy rollouts (384k transitions per iteration), and 200 iterations, with fresh randomized plants and noise every rollout. Observations comprise 17 plant-side features (log energies of error, disturbance proxy, and reference, fast/slow EMA trend, error–filtered-reference correlation, filter norm and update norm, previous action, and time) plus 8 gate-side features (last-interval-was-probe flag, current probe rate $1/K_t$, backoff level, cooloff, normalized wealths, last paired statistic).

The Meta-AF–style baseline is a GRU (64 hidden units) over the 17 plant-side features producing the same $(\log_{10}\mu, \lambda)$ actions, trained by truncated BPTT (25 intervals) through the differentiable simulator on the identical training distribution, minimizing mean log error energy, 60 episode-batches of 256 plants, Adam 10^{-3}. EGA-GRU uses the identical architecture and training recipe but observes all 25 features (including gate state) and is trained *through* the gate: probe, rollback, commit, and demotion events are treated as nondifferentiable environment dynamics, anchor snapshots are detached from the autograd graph, and gradients flow through the FxNLMS updates on live intervals only.

B.7 COMPUTE

All experiments run on a single NVIDIA RTX A6000. The vectorized environment executes $\approx$ 58k control intervals/s at 1024 parallel plants. PPO training takes $\approx$ 1 h per seed, the GRU baseline $\approx$ 25 min, and the full evaluation suite $\approx$ 45 min. Total compute for all reported results is under 12 GPU-hours.

C Ablations and Additional Results

Table 2: Polarity-fault suite ($\hat{S}$ sign-inverted, 512 plants). Two net-harm flags are reported side by side: harm_{05} ($\sum e^2 > 1.05 \sum d^2$, the flag quoted in the abstract and Sec. 6) and the stricter harm_{25} ($\sum e^2 > 1.25 \sum d^2$). The blind adapters amplify the error. All supervised arms cap the fault at similar raw decibels, and on medians EWMA (-1.6 dB) is slightly ahead of our reflex tier (-1.8). The difference shows in the harm flags, not in raw dB. The watchdog/CUSUM/EWMA rows end 99.8/100.0/100.0% of episodes net-harmful at harm_{05}, vs. 93.0% for ours. Their fixed sensitivities tolerate sub-threshold amplification. The EGA caps are bounded but not zero (budget, overheads and reflexes). Headline entries are quoted in Sec. 6. †our configuration.

Method	med (dB)	p10 (dB)	harm_{05}%	harm_{25}%
FxNLMS (best μ)	-7.5	-12.8	100.0	99.6
VSS-FxNLMS	-5.0	-10.0	100.0	99.6
Hysteresis freeze/adapt	-7.0	-12.3	100.0	99.6
FxNLMS μ=0.02 + online $\hat{S}$	0.9	-2.0	30.3	18.9
GRU adapter	-3.9	-7.4	99.8	99.2
PPO (no gate)	-6.4	-11.1	100.0	99.4
Watchdog + GRU (post hoc)	-1.9	-3.6	100.0	88.3
Reflexes-only + GRU (post hoc)	-1.2	-3.4	93.4	55.9
EGA + GRU (wrapped post hoc)	-1.2	-3.3	93.0	55.5
Watchdog (matched ckpt)	-2.4	-4.0	99.8	93.6
CUSUM (matched, null-calib.)	-1.8	-3.4	100.0	92.0
EWMA (matched, null-calib.)	-1.6	-3.1	100.0	83.4
EGA no-VP (matched)	-1.5	-3.0	93.2	67.4
EGA-RL (ours)†	-1.4	-3.4	89.1	60.0
EGA-Full (ours)†	-1.8	-3.5	93.2	71.3
EGA-Reflex+VP (ours)†	-1.8	-3.6	93.0	71.5

Eriksson online-$\hat{S}$ re-estimation caps the polarity fault (row above), but the repair is not free: under OOD noise, where the factory calibration is not wrong, its continual re-estimation costs attenuation relative to tier 1 (2.4 vs. 4.8 dB). The fault-specific fix does not carry over to the benign-shift regime EGA is built for.

C.1 Gate-component ablations and adversarial stress

Table 3 reports gate ablations on the main suite (256 held-out plants) with the trained EGA-RL adapter and with an adversarial fixed $\mu = 0.2$ proposal. The trained adapter barely needs the circuit breaker (its features let it avoid divergence), but for adversarial proposals the breaker roughly halves realized harm (random policy: 26.2% vs. 50.8% without it, and fixed μ=0.2: 40.6% vs. 62.5%). Exponential backoff stops the rollback thrashing: without it the adversarial proposal triggers 10.4 rollbacks per episode and its p10 drops from -1.7 to -3.5 dB. Removing demotion moves the trained adapter's harm from 0.8% to 1.2% and the adversarial rate from 40.6% to 37.5%. At $n = 256$ these are differences of one to four episodes, and we report them without a significance claim. The demotions column reports this plainly: with the staleness margin $\gamma_s = 0.25$ set so that typical measured drift alone cannot fire it, demotion *never fires* in these evaluation suites (0.00/episode on fault_mid, and likewise elsewhere). Validation distrust and the safe-side refuge re-ground the system first (Sec. 6). Demotion stays in the design as untriggered insurance for the hazard class it guards (a committed anchor plus a plant change that the refuge criteria miss), and we credit it with no measured win. Smaller probe periods K pay probe overhead (K=4: median 4.4 dB with more probe-transient harm flags). With larger K the gain test runs short of probes: commits fall from 0.30 (K=4) to 0.34 (K=32) per episode, since there are too few probes to certify gains within an episode, and detection slows. This is the probe-rate versus latency trade of the betting design.

Table 3: Gate ablations (main suite). rb/cm/dm = rollbacks/commits/demotions per episode.

Configuration	med (dB)	p10 (dB)	harm%	rb	cm	dm
Full EGA-RL	4.4	1.2	0.8	0.8	0.4	0.00
– circuit breaker	4.6	2.0	1.2	0.0	0.4	0.00
– demotion (stale anchors kept)	4.7	1.6	1.2	0.5	0.5	0.00
– exponential backoff	4.4	1.3	1.6	1.3	0.3	0.00
probe period K=4	4.4	1.4	1.6	0.7	0.3	0.00
probe period K=16	4.4	0.9	2.3	0.6	0.4	0.00
probe period K=32	4.4	0.6	2.3	0.5	0.3	0.00
Adversarial μ=0.2 + gate	0.0	-1.7	40.6	3.9	0.1	0.00
– circuit breaker	-1.4	-8.7	62.5	1.7	0.1	0.00
– demotion	0.2	-1.3	37.5	3.9	0.1	0.00
– backoff	-0.0	-3.5	47.7	10.4	0.1	0.00
Adversarial μ=0.2 + watchdog	0.3	-2.1	40.6	3.6	0.0	0.00
Random policy + gate	0.3	-0.8	26.2	4.0	0.1	0.00
Random policy + watchdog	0.5	-1.5	34.8	3.7	0.0	0.00

C.2 CERTIFICATE-BUDGET SWEEP

Across $\delta \in \{0.01, 0.05, 0.2\}$, realized net-harm episode rates for the trained adapter are 0.8%, 1.2%, 2.3% respectively, flat in δ. The reading matters here: δ bounds the probability that any *certificate* is violated (a test firing falsely or failing its coverage), not the frequency of net-harm episodes, which is dominated by many small, bounded, in-budget excursions. That flatness, together with the conservatism of the union bounds, is the expected signature. Adversarial proposals (random, fixed μ=0.2) show the same pattern, with severe harm ($> 25\%$ excess energy) far rarer than the 5%-threshold rate (raw distributions included in the supplementary material).

C.3 THE DUTY-CYCLING ADVERSARY

The adversary that motivates the theory confines harm to intervals the monitor does not observe. We realize it as a proposal that spikes $\mu = 0.3$ for a single interval every fourth interval and hides at $\mu = 10^{-4}$ otherwise, which defeats the breaker's two-interval patience by construction (512 held-out plants). Ungated, it produces severe harm ($> 25\%$ excess energy) in 70.9% of episodes (median -3.5 dB). The fixed-schedule watchdog still catches part of the damage (38.3% severe harm), because in this plant class harm injected through the update rule *persists in the filter state* between spikes. This is a structural mitigation that we did not assume in the theory. The null-calibrated charts do *not* improve on it (EWMA 40.2% and CUSUM 45.6% severe harm, p10 -2.9/-3.6 dB). CUSUM is the *worst* supervisor here, because its reference subtraction $S_j = \max(0, S_{j-1} + \bar{Z}_j - k)$ drains accumulated evidence between spikes, exactly the blind-spot mechanism of Proposition A. The *reflexes-only* control (randomized probes, virtual-harm reflex, breaker, backoff, e-tests off) cuts severe harm to 12.3% (median -0.1, p10 -1.1 dB), 3.1–3.7× below every fair classical supervisor at matched false-alarm budget and matched rollback rates (4.1–4.4 rollbacks/ep. across all supervised arms). The charts' statistics drain or reset between spikes and the breaker's patience is defeated by construction, while the virtual-harm reflex fires on *per-interval* virtual evidence and backoff escalates the response. That is protection no fixed schedule or drained chart provides. Full EGA is statistically indistinguishable from the reflex tier on this adversary, at 12.2% severe (median -0.2 dB, p10 -1.1 vs. -2.6 dB for the watchdog), with overlapping bootstrap CIs $[10.5, 13.8]\%$ vs. $[10.7, 14.0]\%$. The harm e-process adds no measurable severe-harm reduction beyond the reflex tier. What it still adds on this adversary is what it adds everywhere: per-trial commit certificates, the tunable δ-budget, demotion semantics and the audit ledger, priced at +0.92 dB paired (Sec. 6). At the mild 5% threshold the tiers likewise coincide (45.6% vs. 45.6%), both below the watchdog's 54.4%. Each duty row pools four seed batches of 128 plants *times three gate seeds* (offsets $+0, 7, 14$ applied to every base seed, 1536 episodes in total per row), so the pooled figures average over gate-coin variability.

Checking for an evidence-access confound. The reflex tier's virtual-harm reflex sees a per-interval reconstruction $\hat{e}(a)$ that the charts above do not, so the win above might in principle be an unfair comparison rather than a mechanism difference. We added two rows to check. The first row feeds CUSUM/EWMA the identical per-interval z_v stream the reflex tier's vp reflex bets on (`charts_use_virtual=True`, an extra $S_j \mathrel{+}= z_v - k$ / EMA update every non-probe inter-

val), null-calibrated on the stationary suite to the watchdog's false-alarm budget. An initial grid ($h \in \{16, 24, 32, 48, 64\}$, $\tau \in \{0.5, 0.7, 1.0, 1.4, 2.0\}$, `calibrate_vp_charts.py`) undershot the budget by an order of magnitude everywhere. The extra per-interval drain empties the statistic toward its floor twice as often under the null, so matching the budget needs thresholds *below* the non-VP defaults, not above them. Re-sweeping downward (`calibrate_vp_charts_v2.py`, $h \in \{1, 2, 3, 4, 6, 8, 12, 16\}$, $\tau \in \{0.10, 0.14, 0.18, 0.22, 0.28, 0.35, 0.5\}$) gives the correct closest-from-above pair: $h_{\text{vp}} = 2.0$ (0.057 rollbacks/ep) and $\tau_{\text{vp}} = 0.10$ (0.065 rollbacks/ep) against the watchdog's 0.047 budget. The overlap points (h=16, τ=0.5) reproduce the original sweep's rate exactly. CUSUM+VP and EWMA+VP reach 21.4/21.2% severe harm, roughly *half* the non-VP charts' 45.6/40.2%, but still $\approx 1.7\times$ above the full reflex tier. The second row strips the vp reflex entirely (`use_virtual=False` with `use_etests=False`), leaving only randomized real-probe reflexes and backoff. Severe harm rises to 25.5% (CI $[23.3, 27.7]$), *worse than either virtually-fed chart*, not just the full reflex tier. Properly calibrated, virtual evidence is therefore the larger contributor. Wherever it is fed it roughly halves severe harm, in the reflex tier's own channel ($25.5\% \to 12.3\%$) as much as in the classical charts ($45.6\% \to 21.4\%$, $40.2\% \to 21.2\%$). The unpredictable coins, the escalating backoff and trust gating account for the *remaining* $\approx 1.7\times$ edge of the full reflex tier over the best virtually-fed chart. Either way the result is not an evidence-access artifact: no chart, fed virtually or not, matches the reflex tier's 12.3%. Unlike the e-process's *analytic* $\log(1/\delta_k)$, the non-VP charts' thresholds required an offline null-calibration sweep. We ran it on the stationary suite to the watchdog's false-alarm rate (the closest CUSUM grid point lands slightly *above* it, generous to CUSUM). Such a sweep is unavailable on-device, where the null law drifts per user, fit and session, and the +VP charts above needed a second, independent sweep at a different scale for the same reason. The certified tier's sized budget on realized harm is validated in Fig. 3.

C.4 Additional figures

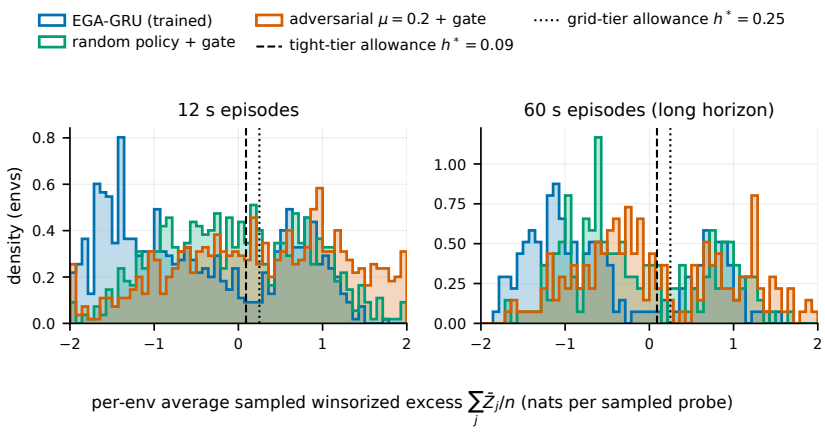


Figure 3: The certified tier's ledger audit: per-plant sampled winsorized excess vs. budget (main/polarity/fault_mid suites, 12/60 s horizons). The positive mode is the caught polarity fault. The per-cell time-average stays below the reference rate (max 0.13/0.04 nats/interval).

Sampled averages behind Fig. 3 (main text, 12 s/60 s): EGA-GRU gives -1.23/-1.08 on main, -0.91/-1.05 on fault_mid, and 0.72/0.71 under the polarity fault. An adversarial FxNLMS μ=0.2 proposal gives -0.07/-0.36, -0.08/-0.41 and 0.93/1.01. A random proposal gives -0.39/-0.57, -0.41/-0.62 and 0.73/0.77. Per plant, which is the form the ledger audit of Prop. 3 reports, 0.0% of plants exceed the reference rate on main and 0.8% on fault_mid, against 69.5% under the polarity fault, falling to 1.6% at 60 s. Those excursions land in the ledger's run-threshold terms (event-terminated runs), not in its allowance rate.

C.5 Monitor separation: the synthetic comparison

Proposition A concerns monitors, so we compare them on a common statistic stream with no ANC loop in the way (`code/experiments/separation_v2.py`,

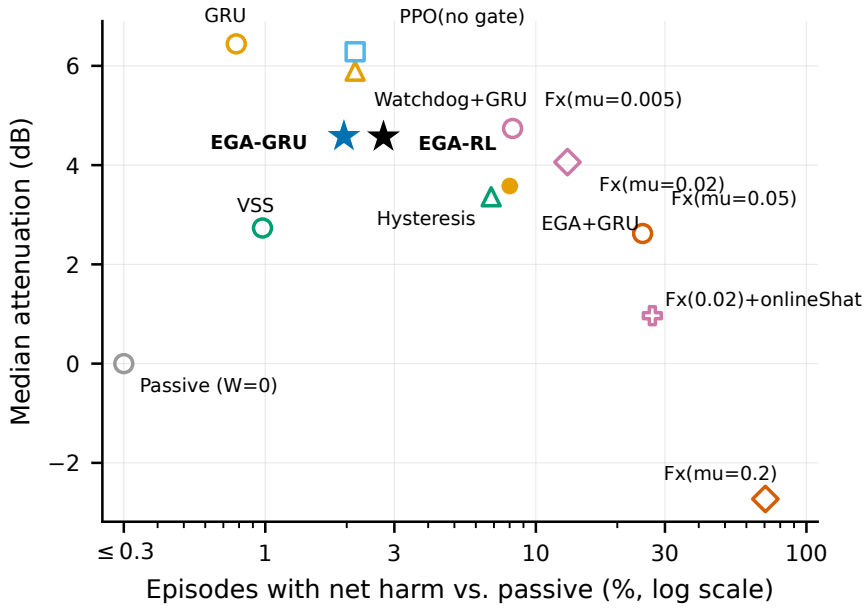


Figure 4: Safety–performance frontier on the main suite (open markers denote uncertified systems, filled markers denote EGA-wrapped systems). Classical baselines trade attenuation against net-harm probability along a frontier that certified systems break.

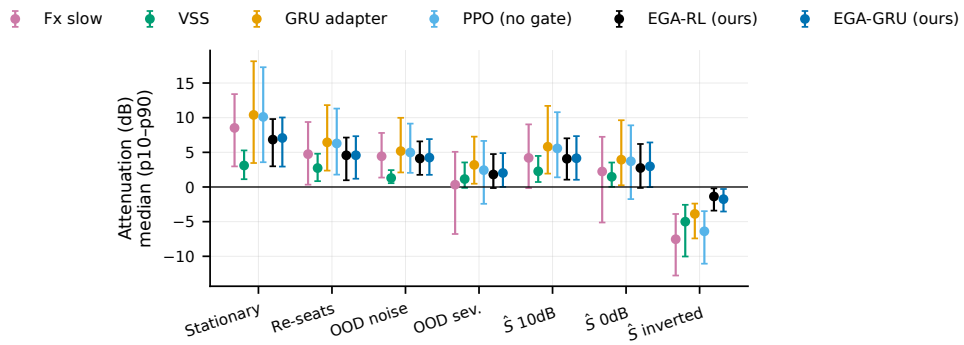


Figure 5: Median (p10–p90) attenuation across the evaluation suites. Uncertified learned adapters are strong under moderate shift but fail up to the saturation cap under calibration faults (rightmost). EGA systems are competitive in distribution and remain near passive under every fault.

`results/separation_v2.json`). The null residuals are *empirical*. Every per-probe block-mean $\bar{Z}_j$ the harm family produced during real gated episodes (EGA-GRU on the main suite, $n \approx 1.8 \times 10^4$ draws) is logged, the pool is centred, and we sample from it, so the fluctuation law is the deployed one (clipped, with an atom at $\pm c$, far from Gaussian, rms 1.26 nats). The adversary adds a constant conditional harm h and the sum is re-clipped at $c = 2$. Horizon $T = 20{,}000$ probes, which at $K = 8$ and 16 ms intervals is 43 minutes (0.71 h) of listening, with 40 repetitions per level.

Every monitor is calibrated by bisection *under the null* to the same false-alarm probability $\delta/3$ over the horizon, so no monitor is handicapped by its tuning. The memoryless (Shewhart) watchdog with patience 2 gets $\tau = 2.00$. Page's CUSUM $S_j = \max(0, S_{j-1} + \bar{Z}_j - k)$ with reference $k = 0.25$ gets $b = 18.7$. The EWMA with $\alpha = 0.05$ gets $\tau = 0.65$. EGA's mixture over $\mathcal{H}_{\text{harm}}$ uses its *analytic* threshold $\log(1/\delta_k)$ and needs no calibration at all. That is the practical point of the e-process route, since on-device the null law changes with the user, the fit and the session. Median detection time (in probes, where T means the median run never fired) and cumulative harm $h \cdot \mathbb{E}[\min(N, T)]$ (nats):

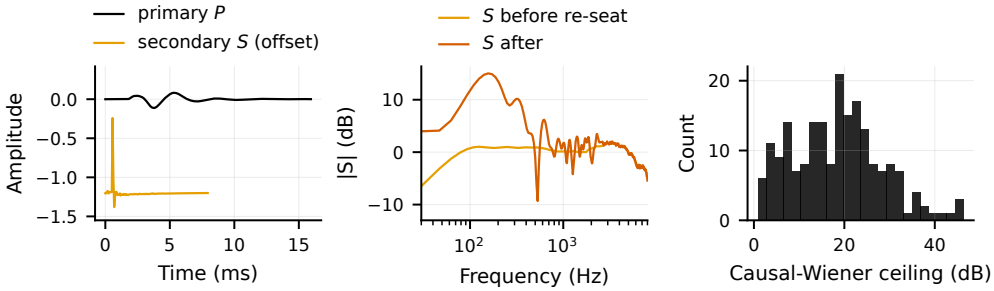


Figure 6: Left: example primary/secondary impulse responses. Middle: secondary path magnitude before/after a severity-0.5 re-seat. Right: distribution of the causal-Wiener attenuation ceiling across 200 sampled plants.

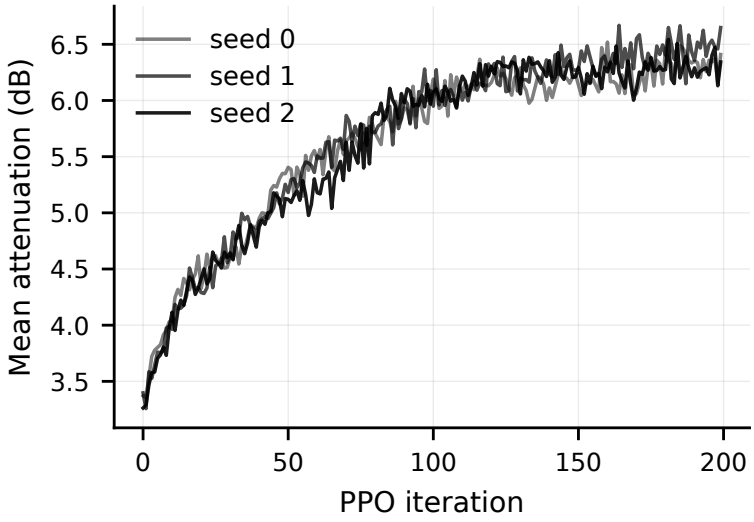


Figure 7: PPO training curves (mean attenuation per iteration, 3 seeds, final training attenuation $6.4/6.7/6.3$ dB).

h	0.02	0.05	0.1	0.2	0.4	0.8
mixture: detect / harm	20000 / 389	3872 / 216	1086 / 109	202 / 44	70 / 30	26 / 22
CUSUM k=0.25	20000 / 387	20000 / 940	20000 / 1441	1368 / 385	106 / 50	36 / 30
EWMA α=0.05	20000 / 388	20000 / 928	20000 / 1465	3344 / 1054	190 / 117	32 / 31
Shewhart τ=2.00	20000 / 400	20000 / 1000	20000 / 2000	20000 / 4000	20000 / 8000	20000 / 16000

Worst case over h: mixture reaches 389 nats $= 0.0194\,T$. CUSUM reaches 1441 ($0.0720\,T$, $3.7\times$ the mixture). EWMA reaches 1465 ($0.0733\,T$, $3.8\times$). Shewhart reaches 16000 ($0.8000\,T$, $41.2\times$).

Scaling exponents (descriptive). Repeating the whole comparison at $T = 4000$ and fitting $\text{Harm} \sim T^p$ across the two horizons (`results/scaling.json`) gives

$$p_{\text{mixture}} = 0.53, \quad p_{\text{EWMA}} = 0.73, \quad p_{\text{CUSUM}} = 0.86, \quad p_{\text{Shewhart}} = 1.00.$$

The ordering matches the blind levels of Proposition A, but the exponents are fitted from two horizons only and we do not treat them as asymptotic rates. Shewhart sits at exactly linear because its validity-calibrated $\tau = 2.00$ *equals* the winsorization level $c = 2$. A memoryless threshold on a clipped statistic whose law has an atom at the clip cannot both keep a δ false-alarm rate and ever fire, so its blind level covers the entire range and it accrues hT at every h. CUSUM and EWMA sit in between. Proposition A(a) suggests $\Theta(T\log(1/\delta)/\log(T/\delta))$, whose fitted exponent between these two horizons is $1 - \log\big(\frac{\log(T_2/\delta)}{\log(T_1/\delta)}\big)/\log\frac{T_2}{T_1} = 0.92$ at $\delta = 1/60$, against 0.86 and 0.73 measured. The mixture's 0.53 must *not* be read as an asymptotic $\sqrt{T}$: its worst case over h is attained at the smallest

tested level $h = 0.02$, where the grid floor $2\bar{v}\eta_{\min}T \approx 0.018\,T$ of Proposition A(a5) limits it. The fitted exponent is pre-asymptotic, and a horizon-tuned chart would match the rate (Proposition A(b)). We also read nothing into the CUSUM/EWMA gap: their worst-case harms at $T = 20000$ (1441 vs. 1465 nats) are within 2% of each other, and the exponent difference reflects their different behaviour at the smaller horizon, not a separation between them.

Limits of the comparison. At $h \leq 0.05$ *no* monitor, including ours, detects within 43 minutes. All four accrue $\approx hT$, and the mixture's 389 nats at $h = 0.02$ is essentially that ($hT = 400$). The mixture's advantage is concentrated in $h \in [0.05, 0.4]$ (levels large enough to be detectable in the horizon and small enough to sit inside the fixed-sensitivity charts' blind spots), where it detects in 3872/1086/202 probes at $h = 0.05/0.1/0.2$ while CUSUM and EWMA need 1368/3344 probes even at $h = 0.2$ and never fire in the median run below it. Above $h = 0.4$ all three sequential monitors are comparable. Finally, this is a monitoring experiment: it isolates Proposition A's setting deliberately, and it is not an end-to-end ANC demonstration (see the negative result below).

Long-horizon ANC: a negative result, and a second failure mode of thresholds. We ran 600 s episodes (24 plants) with a persistent sub-threshold adapter (FxNLMS at $\mu = 0.05$) intended to exhibit the separation end-to-end, and found no separation. Every arm accumulates *negative* cumulative excess against passive (EGA -0.24, watchdog -0.32 of the trivial cap), because excess misadjustment makes the candidate worse than a converged *anchor* while it remains far better than *passive*. So nothing accumulates against passive for either monitor to detect. We report this rather than engineering a fault that only EGA survives.

The same runs show a second effect of threshold monitoring, on intervention rate, that complements the blind spot of Proposition A. Attenuation and rollback counts point in opposite directions, so both are reported. On attenuation the watchdog ends *higher*, at 2.10 dB median ($[1.01, 3.31]$) vs. EGA's 1.03 ($[0.60, 1.95]$), with overlapping CIs at $n = 24$ and a 1.07 dB median gap in the watchdog's favor (reflexes-only 1.51 dB). On interventions it fires 69.50 rollbacks per episode against EGA's 34.29, $2.0\times$ as many. A fixed threshold compares one instantaneous statistic to a constant, so it is blind to persistent harm below τ (Proposition A) and *trigger-happy* to fluctuations above it. A sequential test accumulates evidence and paces its firing rate, at a cost here, since EGA's rarer rollbacks carry cooloffs that also freeze benign adaptation. On raw decibels in the benign long-horizon regime the watchdog is still ahead of the *full* tier (Watchdog+GRU 5.19, CUSUM 4.96, EWMA 5.16 vs. EGA-Full 2.38 dB). The gap is what cooloffs, trials and probing cost over ten minutes. The reflex+VP tier closes it to 0.66 dB (Sec. 6). Meanwhile EGA accumulates 0.75 certified commits per episode, while the watchdog, whose anchor is pinned at passive, records 0.00 by construction and can never improve at all. The realized variation measured in these runs (EGA 1.66 benign, 1.88 under the sub-threshold adapter) is consistent with the 1.81 used throughout.