AUDITING EARLY STOPPING AFTER DEPLOYMENT: SEQUENTIAL INFERENCE WITH SELECTIVE FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Changes after deployment can invalidate rules that stop computation or measurement once intermediate results appear sufficient. Early stopping then hides the completed output needed to check its decisions. We study disagreement between early and completed outputs, and show that some violations are statistically indistinguishable from deployments meeting the risk target using the remaining observations alone. Randomised completion restores feedback on selected early stops; under our protocol, it also replaces their proposed outputs at additional cost. We develop sequential inference for this coupling of selective observation and intervention. An information lower bound shows how audit frequency limits detection speed when pre-audit observations reveal no change. A likelihood-ratio alarm matches its inverse audit-frequency dependence under uniform auditing over the analysed risk range, with grid, restart and overshoot penalties. Correcting for selective observation yields anytime-valid upper bounds on average disagreement risk since deployment began under changing distributions. Alarms separately control deployment-wide false-alarm probability when conditional risk always meets its target. Experiments across magnetic resonance imaging, language-model sampling and image classification distinguish observable change from excess disagreement and examine whether returning the reference improves task outcomes. In the primary classifier comparison, adaptive auditing achieves shorter detection delay and lower delivered disagreement at a common normal-operation cost target, with higher post-change expenditure.

1 INTRODUCTION

Early stopping returns an output before completing all allotted computation or measurements. Magnetic resonance imaging (MRI) may stop repeated measurements once an image appears stable. A large language model (LLM) can sample several answers and return the most frequent one, stopping when further samples seem unlikely to change that choice (Aggarwal et al., 2023). An early-exit classifier returns an intermediate prediction without running every network stage (Jazbec et al., 2024).

Before deployment, stopping rules can be calibrated on fully completed examples to meet a risk target. Changes in data, model configuration or implementation can invalidate that guarantee. Our target concerns agreement with the *full-budget reference*, the output obtained by completing all allotted work. *Disagreement risk* is the probability that the proposed output differs from this reference beyond a specified tolerance.

Early stopping then obstructs verification: it omits the reference needed to check its own decisions. Cases that do not stop early have zero disagreement by construction, so recorded losses can remain zero while risk rises. Proposition 1 establishes a stronger obstruction: two deployments can have identical observed-data distributions yet lie on opposite sides of the risk target. No test using those observations can distinguish them. Additional feedback is needed.

An *audit* supplies this feedback by randomly selecting a proposed early stop, recording its proposed output, and completing the remaining work. Known selection probabilities permit inference from selectively observed full outcomes.

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.

Under our protocol, an audit is also an intervention: the user receives the full-budget output instead of the proposed early output. Audit allocation therefore links detection speed, the disagreement actually delivered to users, and completion cost.

Contributions. **(1)** We prove that some risk violations are indistinguishable from within-target deployments without additional feedback, and formalise a protocol coupling observation to delivery (Proposition 1, §4).

(2) We derive an information lower bound for shifts invisible before auditing and construct a likelihood-ratio (LR) alarm with matching inverse audit-frequency dependence under uniform auditing over the analysed risk range, accounting for grid, restart and overshoot penalties (Proposition 2, Theorem 1). Selective-feedback correction also yields cumulative-risk bounds valid under changing distributions (Theorem 2). Deployment-wide false-alarm control is preserved under predictable audit allocation (§4.3–4.5).

(3) Experiments across MRI, LLM sampling and image classification distinguish observable change from excess disagreement and reference agreement from task quality. A controlled classifier study compares detection, delivered disagreement and post-change expenditure at a common normal-operation cost target (§5).

2 Related work

Certified early stopping. Certifying a stopping decision does not establish continued validity after deployment changes. Learn-then-Test and risk-controlling prediction sets provide high-probability risk guarantees from calibration data (Angelopoulos et al., 2025; Bates et al., 2021); conformal risk control bounds expected risk under exchangeability (Angelopoulos et al., 2024). Applications include early-exit classification (Jazbec et al., 2024), uncertainty calibration for MRI measurements (Fischer et al., 2025), image-specific acquisition times (Yang & Zhang, 2026), and LLM compute allocation (Wang et al., 2026). LLM guarantees also concern the answer mode (Cordero-Encinar & Duncan, 2025) or full-budget vote given true switching probabilities, estimated in practice (Chen et al., 2026). We study continued verification when deployment changes can invalidate a guarantee and early stopping hides the output needed to check it.

Sequential risk monitoring. Confidence sequences provide time-uniform bounds (Howard et al., 2021; Waudby-Smith & Ramdas, 2024); risk monitoring uses arriving labels (Podkopaev & Ramdas, 2022) or labels with predictions (Angelopoulos et al., 2023; Zhang et al., 2026). Anytime-valid conformal risk control addresses growing calibration samples (Hultberg et al., 2026). Here, early stopping prevents the full-budget output needed to evaluate disagreement from being produced. Label-free shift tests (Rabanser et al., 2019) and sequential change detectors (Shekhar & Ramdas, 2023; Shin et al., 2024) can monitor inputs or scores, but observable change need not imply excess disagreement, and some violations leave these observations unchanged.

Selective feedback and adaptive observation. Selective-label research studies decision-dependent missing outcomes (Lakkaraju et al., 2017). Active statistical inference handles adaptive queries; we use its augmented inverse-propensity estimator (Zrnić & Candès, 2024). On-off observation control trades detection delay against observation cost (Banerjee & Veeravalli, 2012; 2013); RouteCert certifies risk conditional on the terminal input pattern under adaptive acquisition (Baghi, 2026). Our protocol couples feedback acquisition to output delivery: completion incurs the remaining cost and replaces the proposed output. We analyse the resulting monitoring and delivery trade-off under a deployment-wide false-alarm probability constraint.

3 Problem setup and missing feedback

Setup. For subject i, a system observes $x_{i,1}, \ldots, x_{i,R_{\max}}$ in stages. A rule π_λ inspects prefixes $x_{i,1:k}$ and proposes a delivery stage $d_i(\lambda) \in \{k_{\min}, \ldots, R_{\max}\}$ before auditing, where $k_{\min}$ is the earliest allowed exit. Write $y_{i,d}$ for the stage-d output; $y_{i,R_{\max}}$ is the full-budget output serving as the reference. Nominally $y_{i,d} = \mathcal{D}(x_{i,1:d})$ for a frozen processor $\mathcal{D}$; a delivery fault may alter the

recorded early output. For discrepancy ρ with $\rho(y, y) = 0$ and tolerance $m \geq 0$, define

$$L_i(\lambda) = \mathbf{1}\left[\rho\left(y_{i,d_i(\lambda)}, y_{i,R_{\max}}\right) > m\right], \qquad L_i(\lambda) = 0 \text{ if } d_i(\lambda) = R_{\max}. \tag{1}$$

Candidates $\Lambda = \{\lambda^{(0)} \prec \cdots \prec \lambda^{(J)}\}$ have a fixed order, starting with the never-stop rule $\lambda^{(0)}$. The initial rule may be selected by offline calibration on completed subjects.

Risk targets and scope. Let $\mathcal{F}_{i-1}$ contain the observed history before subject i. The deployed rule $\lambda_i \in \Lambda$ is $\mathcal{F}_{i-1}$-measurable. Define

$$R_i(\lambda) = \mathbb{E}[L_i(\lambda) \mid \mathcal{F}_{i-1}], \qquad \bar{R}_t^{\mathrm{dep}} = \frac{1}{t}\sum_{i=1}^{t} R_i(\lambda_i).$$

These are the *conditional disagreement risk* and *cumulative average disagreement risk* of the proposed rules, before audit corrections. Fix a target $\alpha \in (0, 1)$ and error level $\delta \in (0, 1)$. For exact label disagreement, $R_i(\lambda_i)$ also bounds conditional excess task error relative to the reference, without assuming the reference is correct (App. A). It does not bound absolute task error. MRI instead uses tolerance exceedance in Eq. (1). A *reference failure* makes the full-budget output worse on average than the proposed output under an external task-quality measure; disagreement alone cannot identify which output is better.

Self-censoring. Without audits, early stops hide the full-budget output, while natural completions have zero loss. All observed losses can therefore be zero while risk is high; if no subject completes, none are observed. More strongly, some violations cannot be distinguished from deployments meeting the target.

Proposition 1 (Impossibility without audits)**.** *Condition on a common initial observed history and fix the same history-dependent deployment policy under two data-generating laws P, P'. Let the observed data be $O_i = (x_{i,1:d_i(\lambda_i)}, d_i(\lambda_i), y_{i,d_i(\lambda_i)})$. Suppose P, P' induce the same conditional law of O_i given each observed history, but differ in the unobserved suffix law on stopped subjects. Then $(O_i)_{i \leq t}$ has the same distribution under both laws for every t. For any fixed t, if $\bar{R}_t^{\mathrm{dep}} \leq \alpha$ P-almost surely and $\bar{R}_t^{\mathrm{dep}} > \alpha$ P'-almost surely, with conditional risks evaluated under their respective laws, every test based on $(O_i)_{i \leq t}$, including a randomised test, has the same rejection probability under both laws.*

Prefix monitors and trusted delivery-integrity checks can detect changes in their observables. Proposition 1 concerns laws with identical observed-data distributions; distinguishing them requires additional information. We obtain it through randomised completion.

4 Audited monitoring and intervention

4.1 Randomised audits

Let $\mathcal{H}_i = \mathcal{F}_{i-1} \vee \sigma(O_i)$ contain the current prefix, proposed stage and output, but not the audit randomisation. At a proposed early stop, record the *proposed early output*, the output otherwise intended for delivery. Choose an $\mathcal{H}_i$-measurable probability $p_i \in [p_{\min}, 1]$, with fixed $p_{\min} > 0$, and draw $A_i \mid \mathcal{H}_i \sim \mathrm{Bernoulli}(p_i)$ independently of the unobserved suffix conditional on $\mathcal{H}_i$. If $A_i = 1$, complete the remaining stages and deliver the full-budget output; otherwise deliver the proposed output. For natural completions, set $A_i = p_i = 1$. Thus A_i records complete-feedback availability, not only paid audits.

When $A_i = 1$, the recorded proposal and completed output reveal $L_i(\lambda_i)$, including faults in early-output delivery. A completed trajectory also reveals losses of fixed candidates whose decisions and outputs can be replayed from it. Delivered disagreement is $L_i^{\mathrm{del}} = (1 - A_i)L_i(\lambda_i)$; conditional independence gives

$$\mathbb{E}[L_i^{\mathrm{del}} \mid \mathcal{H}_i] = (1 - p_i)\mathbb{E}[L_i(\lambda_i) \mid \mathcal{H}_i]. \tag{2}$$

Auditing reveals the proposed rule's loss while removing its delivered disagreement. For a fixed rule, uniform audits independent of its saving retain a fraction $1 - p$ of expected savings, without additional audit overhead (Proposition 4).

Alarms test $H_0 : R_i(\lambda_i) \le \alpha$ for every i; confidence bounds quantify $\bar{R}_t^{\mathrm{dep}}$. An alarm neither locates a violation nor certifies current risk. We first analyse an LR alarm selected before each subject (§4.2), then return to prefix-adaptive auditing for cumulative-risk inference (§4.3).

4.2 Audit frequency and detection delay

Write $\mathrm{kl}(a, b) = a \log(a/b) + (1 - a) \log((1 - a)/(1 - b))$, with continuous endpoint limits. When pre-audit observations reveal no change, audit information limits detection speed.

Proposition 2 (Information lower bound for silent shifts)**.** *Call a shift* silent *if it changes the loss law but leaves the pre-audit observation law unchanged. Consider two laws P_0, P_1 under a common initial history and the same deployment and audit policies. Suppose O_i has the same conditional law under both laws given each common observed history, and the rule stops every subject early. Let $L_i \sim \mathrm{Bernoulli}(\alpha)$ under P_0 and $L_i \sim \mathrm{Bernoulli}(r)$ under P_1, where $r \in (\alpha, 1]$, independently of O_i and the past. Assume that the conditional law of the completed suffix given $(\mathcal{H}_i, L_i)$ is the same under both laws, so completion provides no additional information about the change beyond L_i. For any audit scheme satisfying $A_i \perp L_i \mid \mathcal{H}_i$ and $\Pr(A_i = 1 \mid \mathcal{H}_i) \le p \in (0, 1]$, any alarm stopping time τ based on the observed history with $P_0(\tau < \infty) \le \delta$, any $0 \le \beta < 1 - \delta$, and any integer $n \ge 1$,*

$$P_1(\tau \le n) \ge 1 - \beta \quad \Longrightarrow \quad n \ge \frac{\mathrm{kl}(1 - \beta, \delta)}{p\, \mathrm{kl}(r, \alpha)}.$$

If $P_1(\tau < \infty) \ge 1 - \beta$, the same expression lower-bounds $\mathbb{E}_1[\tau]$. This result concerns a change at the start of deployment.

Each audited loss contributes $\mathrm{kl}(r, \alpha)$ nats in this model; the suffix supplies no further information. The bound counts deployment subjects, not audits, and allows history-adaptive auditing.

An LR alarm with matching audit-frequency dependence. The LR sample must be selected independently of the current subject, conditional on the past. Distinguish its inclusion indicator B_i from complete-feedback availability A_i. Draw $B_i \mid \mathcal{F}_{i-1} \sim \mathrm{Bernoulli}(\bar{p}_i)$ with $\mathcal{F}_{i-1}$-measurable $\bar{p}_i$, independently of the subject given $\mathcal{F}_{i-1}$. On early stops, $A_i = B_i$ and $p_i = \bar{p}_i$; on natural completions, $A_i = p_i = 1$. Only $B_i = 1$ subjects enter the LR sample, including zero losses from natural completions. Hence $L_i \mid \mathcal{F}_{i-1}, B_i = 1 \sim \mathrm{Bernoulli}(R_i(\lambda_i))$, where $L_i = L_i(\lambda_i)$; prefix-dependent selection need not preserve this law.

For a fixed grid $\alpha < r_1 < \cdots < r_G < 1$, fixed positive integer restart times s_u, and fixed weights $w_u \ge 0$ with $\sum_u w_u \le 1$, set

$$\mathrm{LR}_g(L) = \left(\frac{r_g}{\alpha}\right)^L \left(\frac{1 - r_g}{1 - \alpha}\right)^{1-L}, \qquad M_t^{\mathrm{LR}} = \sum_u w_u \frac{1}{G} \sum_{g=1}^{G} \prod_{\substack{s_u \le i \le t \\ B_i = 1}} \mathrm{LR}_g(L_i).$$

Empty products equal one, including future restarts. Let $\tau = \inf\{t \ge 1 : M_t^{\mathrm{LR}} \ge 1/\delta\}$.

Theorem 1 (Likelihood-ratio disagreement alarm)**.** *(i) Under the selection scheme above and $H_0 : R_i(\lambda_i) \le \alpha$ for every i, M_t^{LR} is a nonnegative supermartingale with $M_0^{\mathrm{LR}} \le 1$. Consequently, $\Pr(\tau < \infty) \le \delta$, under arbitrary drift satisfying H_0. (ii) Fix a change time s. For every $i \ge s$, suppose $R_i(\lambda_i) \ge r > \alpha$ and $\bar{p}_i \ge p > 0$ on $\{\tau \ge i\}$. Let $s' \ge s$ be the first restart at or after s, with weight $w' > 0$. For every g such that $I_g = \mathrm{kl}(r, \alpha) - \mathrm{kl}(r, r_g) > 0$, the alarm time satisfies*

$$\mathbb{E}[(\tau - s')^+] \le \frac{\log(G/(\delta w')) + \log(r_g/\alpha)}{p\, I_g}.$$

Here I_g is the expected log-likelihood gain per selected subject at risk r. The delay bound requires $I_g > 0$; it does not apply to every $r > \alpha$ for a fixed alternative grid. The pre-alarm conditions include the subject triggering the alarm; fallback starts with the next subject. Delay from the true change time s adds at most the restart wait $s' - s$. At $\alpha = 0.05$, our grid $\{0.1, 0.2, 0.4, 0.8, 0.95\}$ satisfies $I_g \ge 0.68\, \mathrm{kl}(r, \alpha)$ for $r \ge 0.1$, choosing the largest $r_g \le r$ (App. G). Under uniform auditing at rate p, the bounds thus share inverse dependence on $p\, \mathrm{kl}(r, \alpha)$ over this range. Grid discretisation reduces the information rate to I_g; grid mixing, restart weighting and overshoot contribute the numerator terms. Restart schedules and the positive-drift range appear in App. G.

4.3 RISK ESTIMATION FROM AUDITED FEEDBACK

Prefix-dependent auditing can make the raw average of audited losses unrepresentative of deployment risk. For the deployed rule or a replayable candidate λ', let $f_i(\lambda') \in [0,1]$ be an $\mathcal{H}_i$-measurable loss prediction and define the augmented inverse-propensity weighted (AIPW) increment

$$Z_i(\lambda') = f_i(\lambda') + \frac{A_i}{p_i}\big(L_i(\lambda') - f_i(\lambda')\big). \tag{3}$$

It is observable, since the loss is needed only when $A_i = 1$. Conditional independence and the tower property, using predictability of λ_i, give

$$\mathbb{E}[Z_i(\lambda_i) \mid \mathcal{F}_{i-1}] = R_i(\lambda_i).$$

Thus $Z_i(\lambda_i) - R_i(\lambda_i)$ is a martingale difference with respect to $(\mathcal{F}_i)$ despite changes in the rule and data distribution. The same identity holds for each fixed replayable candidate. Prediction affects variance, not validity. Relative to inverse-propensity weighting (IPW; $f_i = 0$), it reduces the audit-induced variance term when its conditional squared loss error is smaller (Proposition 4).

The increments lie in $[1 - 1/p_{\min}, 1/p_{\min}]$. We rescale them to $[0,1]$ for an empirical-Bernstein confidence sequence (CS). Unlike the prefix-based predictor f_i, its centres c_i use only past information.

Theorem 2 (Anytime cumulative-risk monitoring under self-censoring)**.** *Let λ_i be $\mathcal{F}_{i-1}$-measurable, and let $p_i \in [p_{\min}, 1]$ and $f_i = f_i(\lambda_i) \in [0,1]$ be $\mathcal{H}_i$-measurable, with fixed $p_{\min} > 0$. Suppose $A_i \mid \mathcal{H}_i \sim \mathrm{Bernoulli}(p_i)$, conditionally independently of the unobserved suffix. Let the centres $c_i \in [0,1]$ be $\mathcal{F}_{i-1}$-measurable, and let $\mathcal{V} \subset (0,1)$ be a fixed nonempty finite set. Then, with probability at least $1-\delta$, simultaneously for all $t \geq 1$,*

$$\bar{R}_t^{\mathrm{dep}} \leq U_t^{\mathrm{dep,EB}} := \bar{Z}_t + \omega \min_{\nu \in \mathcal{V}} \frac{\log(|\mathcal{V}|/\delta) + \psi(\nu) V_t}{\nu t},$$

where

$$V_t = \sum_{i \leq t} \left(\frac{Z_i - \ell}{\omega} - c_i \right)^2, \qquad \psi(\nu) = -\log(1-\nu) - \nu,$$

$Z_i = Z_i(\lambda_i)$, $\bar{Z}_t = t^{-1}\sum_{i=1}^t Z_i$, $\ell = 1 - 1/p_{\min}$ and $\omega = 2/p_{\min} - 1$. The bound permits arbitrary drift and updates of λ_i and p_i satisfying these conditions.

Optional betting alarm. The same increments also test H_0 under prefix-adaptive auditing. For fixed restarts s_u, use $K_t^{(u)} = \prod_{s_u \leq i \leq t}[1 + \eta_i(Z_i - \alpha)]$. Bets $\eta_i \geq 0$ and their clipping bounds are $\mathcal{F}_{i-1}$-measurable and keep every factor positive. Under H_0,

$$\mathbb{E}[1 + \eta_i(Z_i - \alpha) \mid \mathcal{F}_{i-1}] = 1 + \eta_i(R_i(\lambda_i) - \alpha) \leq 1.$$

Thus a fixed nonnegative mixture with weights summing to at most one controls deployment-wide false alarms at threshold $1/\delta$ (Ville, 1939). Unlike the LR result, we give no delay bound for this alarm. Corollary 1 supplies full conditions; App. A also constructs the betting confidence sequence used by the optional stationary selector (Waudby-Smith & Ramdas, 2024).

4.4 EVIDENCE-DRIVEN AUDIT ALLOCATION

Evidence can guide spending before reaching an alarm threshold. For $\mathcal{F}_{i-1}$-measurable $e_{i-1} \in [0,1]$ and a nondecreasing map $\phi_i : [0,1] \to [0,1]$ chosen from the past, set

$$\bar{p}_i = p_{\mathrm{floor}} + (p_{\max} - p_{\mathrm{floor}})\phi_i(e_{i-1}), \qquad 0 < p_{\mathrm{floor}} \leq p_{\max} \leq 1.$$

Predictability and the positive floor preserve the preceding guarantees under their remaining assumptions: use $p_{\min} = p_{\mathrm{floor}}$ for estimation and $p = p_{\mathrm{floor}}$ in the LR delay bound (Proposition 3). Monotonicity makes spending respond to evidence; it does not establish a performance improvement.

We use each alarm's current log capital divided by its log threshold, maximise across enabled alarms, and clip to $[0,1]$. Smooth escalation uses $\phi_i(e) = \sqrt{e}$; a two-threshold alternative appears in App. A. Selection is not paid auditing: naturally completed subjects incur no extra completion cost. If S_i indicates a proposed early stop and C_i is its remaining cost, unaffected by the selection coin, expected extra cost is $\mathbb{E}[\bar{p}_i S_i C_i]$ when $B_i \perp (S_i, C_i) \mid \mathcal{F}_{i-1}$. Predictability preserves validity, but does not fix normal-operation expenditure. We calibrate that cost separately in §5.2.

4.5 CLOSED-LOOP DEPLOYMENT

Algorithm 1 combines inference and fallback. Here δ is the alarm-family budget: LR and betting each receive $\delta/2$, so a union bound controls any false alarm by δ. Cumulative-risk bounds and optional certification use separate confidence levels. Once an alarm fires after subject i, full completion begins at $i + 1$ and remains latched.

Algorithm 1 Audited deployment

1: Input: target α, alarm budget δ, confidence levels, initial rule λ_1, audit floor/ceiling, optional fixed replayable candidates Λ.
2: Initialise LR and betting alarms at level $\delta/2$, confidence sequences at their own levels, and fallback flag as false.
3: **for** subject $i = 1, 2, \ldots$ **do**
4: From $\mathcal{F}_{i-1}$ choose $\bar{p}_i$, bets, clipping ranges and centres; independently select $B_i \sim \mathrm{Bernoulli}(\bar{p}_i)$ conditional on $\mathcal{F}_{i-1}$.
5: Run λ_i without using B_i; record its proposed stage/output and prefix-based loss predictions before revealing missing losses.
6: On early stops set $A_i = B_i$, $p_i = \bar{p}_i$; on natural completions set $A_i = p_i = 1$. Complete and deliver the reference if $A_i = 1$, otherwise the proposal.
7: Form Z_i using recorded predictions; update the cumulative bound, betting alarm and any replayable-candidate CSs. Update LR only if $B_i = 1$.
8: Use revealed losses to train future predictions. If either capital reaches $2/\delta$, latch fallback permanently.
9: If latched, set $\lambda_{i+1} = \lambda^{(0)}$; otherwise retain λ_i, or apply the optional stationary selector.
10: **end for**

Predictors need not be calibrated for validity; implementation details and the complete procedure appear in App. A. The optional stationary extension certifies fixed replayable candidates under i.i.d. subjects, selecting the last candidate whose own upper bound and all preceding bounds are at most α (Theorem 3). These extra assumptions are not required for the default frozen-rule monitoring loop.

5 EXPERIMENTS

A system change may be observable before its effect on disagreement is known. We therefore ask: how does change detection differ from detecting excess disagreement; how do audit policies trade detection and delivery against cost at a common normal-operation target; and when is confirmation worth waiting for, given the task quality of the reference?

Domains and protocol. We study an ImageNet-pretrained cascade on Imagenette (Howard, 2019), repeated low-field MRI from M4Raw (Lyu et al., 2023), and Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024) on GSM8K and MATH-500 (Cobbe et al., 2021; Lightman et al., 2024). Full-budget references are ConvNeXt-Base, full MRI acquisition and the 64-sample plurality answer, respectively. The classifier runs ResNet-18, ResNet-50 and ConvNeXt-Base, exiting at max-softmax ≥ 0.7. Dataset splits and other configurations appear in App. G. Unless stated otherwise, $(\alpha, \delta) = (0.05, 0.05)$. Stopping-rule calibration uses a fixed-sequence Clopper–Pearson test (Clopper & Pearson, 1934); §5.2 separately calibrates audit cost. Shift streams have length 3000 (6000 for LLMs), with change at 1500. The audit-free monitor tests mean changes in the first stopping score and proposed exit stage, with 300 burn-in observations, tolerance 0.02 and total level δ. Audit alarms use the latched fallback in §4.5.

Costs include deferrals, audits and fallback, normalised by full completion in the reported period. Classifier costs sum nominal multiply-adds of executed stages, accounting for changed architectures or resolutions. We distinguish proposed-rule disagreement from delivered disagreement and label outcomes as whole-stream or post-change. Classifier studies use 30 paired streams in §5.1 and 100 in §5.2. Replay uncertainty is conditional on fixed data pools, not new population samples. Budget and grid follow-ups were designed before execution (App. F).

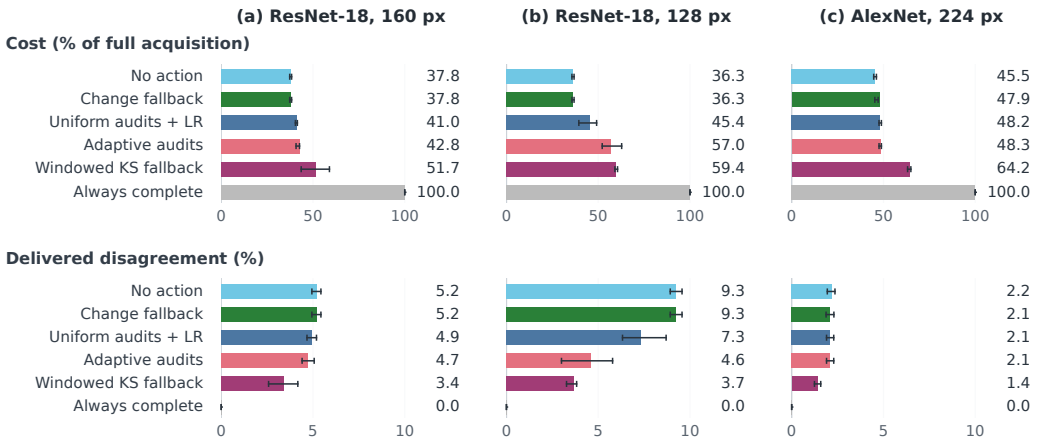

Figure 1: Early-classifier changes: whole-stream cost (top) and delivered disagreement (bottom). Bars: means; whiskers: interquartile ranges over 30 paired streams. Audit rate/floor 0.05, ceiling 0.5; adaptive audits use both alarms. KS: follow-up control. All policies: Table 10.

5.1 Early-branch changes: change versus excess disagreement

We change only the early classifier, holding later stages and the reference fixed. Four configurations were specified before inference: ResNet-18 at 224 (control), 160 and 128 pixels, and AlexNet at 224 pixels. All use the original threshold and split (App. E).

Reducing resolution raises deployment-pool rule disagreement from 0.023 to 0.077 and 0.156, but the two-statistic monitor never alarms in 30 streams. AlexNet replacement has risk 0.020, below target, yet triggers monitor fallback in $8/30$ streams and no audit alarms. A follow-up windowed Kolmogorov–Smirnov (KS) control detects $24/30$ and $30/30$ resolution changes, but also all AlexNet replacements. Thus these changes are not observationally indistinguishable as in Proposition 1: the original misses depend on the monitor. KS tests distributional change under an i.i.d. null; audit alarms test the disagreement-risk null. A change can violate one while satisfying the other.

Audits reveal disagreement, but evidence of excess risk takes time. At 128 pixels, uniform LR detects $16/30$ streams within 1500 post-change subjects; escalation with both alarms detects $27/30$. Whole-stream delivered disagreement falls from 0.093 without intervention to 0.046, and task error from 0.168 to 0.136, at cost 0.570 (Figure 1). Uniform LR costs 0.454 with disagreement 0.073; KS costs 0.594 with disagreement 0.037. At 160 pixels, escalation detects only $5/30$: moderate violations can persist within a finite observation window. On AlexNet replacement, KS fallback costs 0.642 versus 0.483 for those audits, with nearly identical task error. Change detection can prompt intervention, but does not establish a disagreement-target violation.

Component diagnostics. "Both" combines LR and AIPW betting; escalation uses alarm evidence to raise the audit rate, while "joint" also uses the change monitor's evidence. Both wait for an audit alarm before fallback. At uniform rate 0.05, neither AIPW versus IPW nor both alarms versus LR shows a resolved delivered-disagreement improvement: paired 95% bootstrap intervals include zero. Under escalation, both alarms reduce restricted mean delay by 328 subjects versus LR alone, with greater cost; this changes both allocation evidence and the alarm combination (App. E). Prediction reduces estimator variance under resolution changes but increases it under replacement.

Grid sensitivity. At audit rate 0.1, extending the LR grid towards α improves synthetic detection from $75/100$ to $85/100$ at risk 0.1, but only from $0/100$ to $1/100$ at risk 0.06 within 6000 post-change subjects, and can slow larger-shift detection (App. F). Finer alternatives still need sufficient audit information. The audit-policy comparisons above use common rates or floors; we next calibrate normal-operation expenditure.

Table 1: Budget-calibrated policies at 128 pixels, target extra cost 0.03. Means over 100 paired streams. Pre-change extra cost is relative to the unaudited frozen rule; post cost is total post-change cost, each normalised by full completion in its period. Post disagr. is delivered disagreement. "Both" means LR and AIPW betting. Restricted mean delay (RMD) is truncated at 1500 subjects, excluding one pre-change uniform-LR alarm ($n = 99$; 100 otherwise). Costs and disagreement retain all 100 streams.

Policy	Pre-change extra cost	Post cost	Post disagr.	RMD	Detected
Uniform LR	0.034	0.440	0.128	1305	34/100
Smooth LR	0.029	0.507	0.112	1154	63/100
Smooth both	0.030	0.665	0.076	825	90/100
Two-threshold both	0.030	0.597	0.092	974	81/100

5.2 Comparing strategies at a common normal-operation cost target

Using 800 clean calibration images, we calibrate eight policies' audit-rate or floor parameter to extra costs of 0.015, 0.03 (primary) and 0.06 relative to the unaudited frozen rule. Calibration includes false-alarm fallback. We freeze parameters and evaluate 100 paired streams from 3125 disjoint deployment images. Realised pre-change extra costs need not equal the calibration target (Table 1).

At target 0.03, pre-change extra costs are 0.034, 0.029, 0.030 and 0.030 for the four displayed policies. At 128 pixels, smooth escalation with both alarms yields post-change delivered disagreement 0.076, versus 0.128, 0.112 and 0.092 for uniform LR, smooth LR and two-threshold escalation with both alarms. Paired restricted-mean-delay reductions are 477, 329 and 149 subjects. All six Bonferroni-adjusted paired bootstrap intervals exclude zero in the corresponding direction (App. F).

These gains accompany greater post-change cost: 0.665 versus 0.440, 0.507 and 0.597. In this tested setting, smooth escalation with both alarms achieves faster intervention and lower delivered disagreement at the common normal-operation cost target, with higher post-change expenditure. App. F reports all configurations and the lowest-budget IPW calibration that missed its target. No schedule is uniformly best in earlier synthetic comparisons (Tables 28–29); stationary-selector checks appear in App. G.

5.3 When is audit confirmation worth paying for?

We separate waiting for evidence before persistent fallback from deciding whether the reference improves task outcomes. Across 11 conditions and 13 policies, Figures 10–11 report whole-stream medians and interquartile ranges over 10 streams. Representative escalation results use floor 0.05 and ceiling 0.5.

On within-target MRI intensity drift and LLM model replacement, direct change-triggered fallback adds 0.10 and 0.67 of the full budget over no intervention. Alarm-driven escalation adds 0.01–0.05 and triggers no fallback. These savings avoid persistent completion where the measured rule risk remains within target; their task-level value is a separate question.

For the tested 256-token cap, disagreement exceeds the target and direct fallback is competitive: delivered disagreement is 0.002 at cost 0.748, versus 0.005 at cost 0.744 for joint escalation. On three injected bugs missed by the tested monitor, alarm-driven escalation reduces delivered disagreement from 0.35, 0.32 and 0.11 to 0.028, 0.029 and 0.043, while saving 0.16–0.36 of full completion. Trusted integrity checks can also detect these bugs given a trusted processor or output mapping (App. G.1). Audited cases already receive the reference while evidence accumulates; waiting delays persistent fallback, not all intervention.

Does returning the reference improve task outcomes? Under pixelation, the full-budget classifier is less accurate than the proposed early output. Escalation reduces disagreement but increases whole-stream task error from 0.477 to 0.496. Conversely, LLM model replacement worsens task accuracy while disagreement remains 0.043, below target. Direct fallback reduces whole-stream task error from 0.092 to 0.071. Avoiding completion on a within-target change can therefore forgo a task-level benefit. Auditing supplies evidence about excess disagreement; preferring the reference requires separate evidence of task quality.

6 DISCUSSION

Early stopping can hide the feedback needed to verify its decisions. Under our protocol, completion restores that feedback while replacing the proposed output. The information lower bound quantifies the resulting constraint for silent shifts: audit frequency limits detection speed. Experiments identify no uniformly best response. Confirmation can avoid persistent fallback on within-target changes, while direct change-triggered fallback is competitive in the tested token-cap condition. At a common normal-operation cost target, faster intervention accompanies greater post-change expenditure in the primary classifier comparison.

Guarantee scope. The cumulative bound upper-bounds average conditional disagreement since deployment began, before audit corrections. Alarms control the probability of ever falsely rejecting the joint conditional-risk null. The optional stationary selector certifies candidate population risk under additional i.i.d. assumptions. None establishes reference quality: excess disagreement can coexist with a worse reference, and within-target disagreement with degraded task accuracy.

Limitations. Classifier changes are controlled, MRI anomalies are injected, and replay does not add independent subjects. Prospective validation remains needed. Small violations may evade finite-horizon detection; no alarm does not certify current risk below target. The betting alarm has no delay bound. Trusted integrity checks can address the constructed bugs, and optimal audit allocation remains open.

Under our protocol, continued verification is part of the deployment policy: obtaining evidence consumes resources and changes delivered outputs. Deployment must consider together how to obtain evidence, when to commit to full completion, and whether the reference serves the task.

AI USE STATEMENT

We used a generative-AI assistant for literature retrieval, research planning and execution, and drafting and polishing the manuscript. Research assistance included feedback on experimental design, software implementation, result analysis, and review of technical arguments. Verification included reference checks, software tests, and comparisons with experiment outputs. The authors take responsibility for the final manuscript and all reported results.

REPRODUCIBILITY STATEMENT

Appendix A provides the theoretical assumptions, proofs and complete deployment procedure. Appendices C, D, E, F and G describe the data, experimental protocols and additional results. The anonymous supplementary code includes the simulator, inference methods, experiment scripts, fixed seeds, cached trajectories and per-experiment reports, with instructions for reproducing the analyses and figures from the supplied caches. MRI caches contain outputs from the frozen reconstruction pipeline; LLM caches retain extracted answers and generation manifests. The package documents the scope of cache-based reproduction and the dependencies needed for fresh model inference.

ETHICS STATEMENT

Post-market monitoring of high-risk AI systems is addressed in Article 72 of the EU AI Act (European Parliament and Council, 2024). FDA guidance on planned changes to AI-enabled device software (U.S. Food and Drug Administration, 2025) and radiology consensus recommendations (Cuocolo et al., 2025) likewise motivate continued assessment. Our statistical guarantees address disagreement with a reference and do not establish clinical effectiveness or regulatory compliance. M4Raw and the LLM benchmarks are public and de-identified. Audits in a clinical deployment mean scanning a randomly chosen fraction of subjects with the full protocol, which changes scan time and gives those subjects the full protocol's image, whatever its quality; any prospective use would require site-level approval of the audit rate and risk target.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: adaptive-consistency for efficient reasoning and coding with LLMs. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12375–12396, 2023. doi: 10.18653/v1/2023.emnlp-main.761.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnić. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

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

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. arXiv:2110.01052 (2021).

Melika Baghi. Conditional validity for adaptive modality acquisition: when the policy chooses its own calibration group. *arXiv:2608.15520*, 2026. URL https://arxiv.org/abs/2608.15520.

Taposh Banerjee and Venugopal V. Veeravalli. Data-efficient quickest change detection with on-off observation control. *Sequential Analysis*, 31(1):40–77, 2012. doi: 10.1080/07474946.2012.651981.

Taposh Banerjee and Venugopal V. Veeravalli. Data-efficient quickest change detection in minimax settings. *IEEE Transactions on Information Theory*, 59(10):6917–6931, 2013.

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

Wenbo Chen, Puheng Li, Mengyang Liu, Weijie Su, and Tianpei Xie. MARS: margin-adversarial risk-controlled stopping for parallel LLM test-time scaling. *arXiv:2606.12935*, 2026. URL https://arxiv.org/abs/2606.12935.

Charles J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: statistical guarantees and test-time training for reliable reasoning in LLMs. *arXiv:2510.17472*, 2025. URL https://arxiv.org/abs/2510.17472.

Renato Cuocolo, Diana Bernardini, Daniel Pinto dos Santos, Michail E. Klontzas, Tugba Akinci D'Antonoli, Luís Curvo Semedo, Robin Decoster, Merel Huisman, Elmar Kotter, Luis Martí-Bonmatí, Costin Minoiu, Emanuele Neri, Konstantin Nikolaou, Maija Radzina, Evis Sala, Susan C. Shelmerdine, Laurens Topff, Michelle C. Williams, and European Society of Radiology. AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. *Insights into Imaging*, 16:275, 2025. doi: 10.1186/s13244-025-02146-8.

European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), article 72: post-market monitoring, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015. doi: 10.1214/EJP.v20-3496.

Paul Fischer, Jan Nikolas Morshuis, Thomas Küstner, and Christian Baumgartner. CUTE-MRI: Conformalized uncertainty-based framework for time-adaptive MRI. *arXiv preprint arXiv:2508.14952*, 2025. URL https://arxiv.org/abs/2508.14952.

J. L. Hodges, Jr. The significance probability of the Smirnov two-sample test. *Arkiv för Matematik*, 3:469–486, 1958. doi: 10.1007/BF02589501.

Jeremy Howard. Imagenette: a smaller subset of 10 easily classified classes from ImageNet. https://github.com/fastai/imagenette, 2019.

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. doi: 10.1214/20-AOS1991.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control. *arXiv:2602.04364*, 2026. URL https://arxiv.org/abs/2602.04364.

Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: early-exiting with risk control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: evaluating algorithmic predictions in the presence of unobservables. In *ACM SIGKDD*, 2017.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X. Wu. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 10(1):264, 2023. doi: 10.1038/s41597-023-02181-4.

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. arXiv:2110.06177.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential changepoint detection via backward confidence sequences. In *International Conference on Machine Learning*, pp. 30908–30930, 2023. URL https://proceedings.mlr.press/v202/shekhar23a.html.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: a nonparametric framework for sequential change detection. *New England Journal of Statistics in Data Science*, 2(2):229–260, 2024. doi: 10.51387/23-NEJSDS51.

U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: final guidance, 2025. Docket FDA-2022-D-2628; issued December 4, 2024, updated August 18, 2025. https://www.fda.gov/media/166704/download.

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

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: risk control for reasoning on a compute budget. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.03814. arXiv:2602.03814.

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. doi: 10.1093/jrsssb/qkad009.

An Yang et al. Qwen2.5-Math technical report: toward mathematical expert model via self-improvement. *arXiv:2409.12122*, 2024.

Jiawei Yang and Yao Zhang. Conformalized rate-adaptive sensing. *arXiv:2607.26887*, 2026. URL https://arxiv.org/abs/2607.26887.

Guangyi Zhang, Yunlong Cai, Guanding Yu, and Osvaldo Simeone. Prediction-powered risk monitoring of deployed models for detecting harmful distribution shifts. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.02229. arXiv:2602.02229.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017.

Tijana Zrnić and Emmanuel J. Candès. Active statistical inference. In *International Conference on Machine Learning (ICML)*, pp. 62993–63010, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

A AUDIT ECONOMICS AND PROOFS

Proposition 3 (Audit escalation)**.** *Let $0 < p_{\mathrm{floor}} \le p_{\max} \le 1$, let $e_{i-1} \in [0,1]$ be $\mathcal{F}_{i-1}$-measurable, and let $\phi_i : [0,1] \to [0,1]$ be nondecreasing and chosen using $\mathcal{F}_{i-1}$. Set*

$$\bar{p}_i = p_{\mathrm{floor}} + (p_{\max} - p_{\mathrm{floor}})\phi_i(e_{i-1}),$$

and use this rate in the selection scheme of §4.2. Under their remaining assumptions, Theorems 2 and 3 and Corollary 1 retain their guarantees with $p_{\min} = p_{\mathrm{floor}}$. Theorem 1 remains valid under its stated conditions; in its delay bound, one may take $p = p_{\mathrm{floor}}$ since $\bar{p}_i \ge p_{\mathrm{floor}}$. Betting processes may use the predictable per-step range $[1 - 1/\bar{p}_i,\ 1/\bar{p}_i]$. The expected coin-selection rate is

$$\mathbb{E}[B_i] = p_{\mathrm{floor}} + (p_{\max} - p_{\mathrm{floor}})\mathbb{E}[\phi_i(e_{i-1})].$$

Let $S_i = \mathbf{1}\{d_i(\lambda_i) < R_{\max}\}$ indicate a proposed early stop, and let C_i be its potential remaining completion cost, unaffected by the selection coin. With $B_i \perp (S_i, C_i) \mid \mathcal{F}_{i-1}$, the expected paid-audit frequency per subject is

$$\mathbb{E}[B_i S_i] = \mathbb{E}[\bar{p}_i S_i],$$

and the expected additional completion cost is

$$\mathbb{E}[B_i S_i C_i] = \mathbb{E}[\bar{p}_i S_i C_i].$$

We next record the audit-cost and variance identities used in §4:

Proposition 4 (Audit economics)**.** *For a stopped subject, condition on the pre-audit information $\mathcal{H}$ and write $q = \mathbb{E}[L \mid \mathcal{H}]$ and $b = \mathbb{E}[(L-f)^2 \mid \mathcal{H}]$. Under an independent audit coin of probability $p > 0$: (a) $\mathrm{Var}(Z \mid \mathcal{H}) = \mathrm{Var}(L \mid \mathcal{H}) + (1-p)b/p$. Relative to $f = 0$, prediction multiplies the audit-coin variance term by b/q when $q > 0$. (b) For n independent, identically distributed Bernoulli audit losses with no observed failures, the one-sided Clopper–Pearson upper bound is $1 - \delta^{1/n}$. It reaches α when $n \ge \lceil \log\delta / \log(1-\alpha) \rceil$, or 59 at $\alpha = \delta = 0.05$. (c) For a fixed rule, a uniform coin independent of its saving and no additional overhead, the expected saving is $(1-p)\,\mathrm{sav}(\lambda)$, where $\mathrm{sav}(\lambda)$ is its expected saving without audits.*

Proof. Conditionally on $\mathcal{H}$ and L, Z has mean L and variance $(1-p)(L-f)^2/p$; total variance gives (a). Inverting the zero-count binomial probability $(1-r)^n$ gives (b). An unaudited early stop retains its original saving and an audited one retains none, which gives (c). □

Proof of Proposition 1. Starting from the common initial history, equality of the conditional observation laws implies by induction that $(O_i)_{i \le t}$ has the same joint distribution under P and P' for every t. Represent a possibly randomised test by its conditional rejection probability $\varphi_t((O_i)_{i \le t}) \in [0,1]$. Equality of the observation laws gives

$$\mathbb{E}_P[\varphi_t((O_i)_{i \le t})] = \mathbb{E}_{P'}[\varphi_t((O_i)_{i \le t})].$$

For a pair on opposite sides of the risk target at a fixed t, the test's power under P' therefore equals its rejection probability under P.

Such a pair exists even in a two-stage system. Let $R_{\max} = 2$, $k_{\min} = 1$, and $x_{i,1} = 0$. Let the processor return the most recent observation, and let the rule always stop at stage 1. Use exact label disagreement, equivalently $\rho(a, b) = \mathbf{1}\{a \neq b\}$ and $m = 0$. Under P, set $x_{i,2} = 0$; under P', let $x_{i,2}$ be independent Bernoulli variables with mean $r \in (\alpha, 1)$. Every observed record is $O_i = ((0), 1, 0)$ under both laws, whereas $L_i = x_{i,2}$. Consequently, the cumulative average disagreement risk is 0 under P and $r > \alpha$ under P' for every t. □

Lemma 1 (Empirical-Bernstein supermartingale for time-varying means)**.** *Let $X_i \in [0, 1]$ be adapted to $(\mathcal{F}_i)$ with $\mu_i = \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$, let $c_i \in [0, 1]$ be $\mathcal{F}_{i-1}$-measurable, and let $\nu \in [0, 1)$ with $\psi(\nu) = -\log(1 - \nu) - \nu$. Then $M_t = \prod_{i \leq t} \exp\big(\nu(X_i - \mu_i) - \psi(\nu)(X_i - c_i)^2\big)$ is a nonnegative supermartingale with $M_0 = 1$.*

Proof of Lemma 1. For $\xi \geq -1$ and $\nu \in [0, 1)$, $\exp(\nu\xi - \psi(\nu)\xi^2) \leq 1 + \nu\xi$ (Fan et al., 2015, Eq. (4.12), proof of Lemma 4.1). Apply it with $\xi_i = X_i - c_i \in [-1, 1]$ and take conditional expectations: $\mathbb{E}_{i-1} \exp(\nu(X_i - c_i) - \psi(\nu)(X_i - c_i)^2) \leq 1 + \nu(\mu_i - c_i) \leq \exp(\nu(\mu_i - c_i))$. Multiplying by $\exp(-\nu(\mu_i - c_i))$ gives $\mathbb{E}_{i-1} \exp(\nu(X_i - \mu_i) - \psi(\nu)(X_i - c_i)^2) \leq 1$, i.e. the increments of M_t have conditional expectation at most one. □

Proof of Theorem 2. Rescale $\tilde{Z}_i = (Z_i - \ell)/\omega \in [0, 1]$, whose conditional mean is $\tilde{\mu}_i = (R_i(\lambda_i) - \ell)/\omega$. Apply Lemma 1 to $X_i = 1 - \tilde{Z}_i$ (conditional mean $1 - \tilde{\mu}_i$, centre $1 - c_i$): by Ville's inequality (Ville, 1939), for each fixed ν, with probability at least $1 - \delta/|\mathcal{V}|$, for all t, $\sum_{i \leq t} \nu(\tilde{\mu}_i - \tilde{Z}_i) \leq \log(|\mathcal{V}|/\delta) + \psi(\nu)V_t$. A union bound over $\nu \in \mathcal{V}$ and dividing by νt gives $\frac{1}{t}\sum_i \tilde{\mu}_i \leq \bar{\tilde{Z}}_t + \min_{\nu \in \mathcal{V}} \frac{\log(|\mathcal{V}|/\delta) + \psi(\nu)V_t}{\nu t}$; mapping back through the affine rescaling yields the claim. Two information sets are involved and it matters which one each quantity is measurable with respect to. The supermartingale of Lemma 1 is with respect to $(\mathcal{F}_i)$: the bet ν, the centre c_i, the deployed λ_i and the range $[\ell, \ell + \omega]$ are $\mathcal{F}_{i-1}$-measurable, and what the lemma needs is $\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = R_i(\lambda_i)$. That identity is obtained by first conditioning on the larger pre-audit set $\mathcal{H}_i$, with respect to which p_i and f_i are measurable and the audit coin is independent of the suffix: $\mathbb{E}[Z_i \mid \mathcal{H}_i] = f_i + \frac{\mathbb{E}[A_i|\mathcal{H}_i]}{p_i}\big(\mathbb{E}[L_i \mid \mathcal{H}_i] - f_i\big) = \mathbb{E}[L_i \mid \mathcal{H}_i]$, and then applying the tower property. No stationarity or independence between subjects is required beyond these conditional-mean, boundedness and measurability assumptions. A bet or range that depended on the current prefix (an $\mathcal{H}_i$-measurable clipping range, for example) would not generally preserve this supermartingale argument, because the resulting betting parameter could no longer be taken outside the conditional expectation given $\mathcal{F}_{i-1}$. □

Predictable ranges for stationary betting confidence sequences. Betting parameters and clipping ranges are chosen before observing the current subject. If the deployed rule is $\lambda^{(0)}$, every subject is fully observed and candidate increments lie in $[0, 1]$. Otherwise the global audit floor gives $[1 - 1/p_{\min},\ 1/p_{\min}]$. These predictable ranges permit tighter bets during full-acquisition cold starts. Observing that the current subject completes naturally does not by itself justify tightening a bet chosen for its pre-subject conditional risk. Under the scheduled selection scheme, the predictable rate $\bar{p}_i$ permits the per-step range $[1 - 1/\bar{p}_i,\ 1/\bar{p}_i]$.

Theorem 3 (Closed-loop certification, stationary case)**.** *Suppose subjects are i.i.d. so that $R_i(\lambda') = R(\lambda')$ for all i, and that every candidate's output is replayable from a completed trajectory. Fix the candidate order before observing data, with the never-stop rule $\lambda^{(0)}$ first. Let $U_t(\lambda')$ be a level-$(1-\delta)$ upper CS for $R(\lambda')$ built from $Z_i(\lambda')$, and let the deployed candidate be chosen at every step by the fixed-sequence rule $\lambda_{t+1} = \lambda^{(j^\star)}$, $j^\star = \max\{j : U_t(\lambda^{(j')}) \leq \alpha\ \forall j' \leq j\}$ (with $\lambda_{t+1} = \lambda^{(0)}$ if the set is empty). Then $\mathbb{P}\big[\forall t : R(\lambda_{t+1}) \leq \alpha\big] \geq 1 - \delta$. Offline Learn-then-Test uses the special case of the protocol $p_i = 1$ for $i \leq n$ followed by a frozen λ, with a fixed-n test in place of the CS.*

Proof of Theorem 3. If every candidate is safe, the claim is immediate. Otherwise, order the candidates $\lambda^{(0)} \prec \cdots \prec \lambda^{(J)}$ and let $j_{\text{bad}} = \min\{j : R(\lambda^{(j)}) > \alpha\}$ be the first unsafe candidate. The selector deploys an unsafe candidate at some time only if it has reached j_{bad}, which requires

$U_t(\lambda^{(j_{\text{bad}})}) \le \alpha < R(\lambda^{(j_{\text{bad}})})$ at that time. Since $U_t(\lambda^{(j_{\text{bad}})})$ is a level-$(1-\delta)$ CS for $R(\lambda^{(j_{\text{bad}})})$, the event $\{\exists t : U_t(\lambda^{(j_{\text{bad}})}) < R(\lambda^{(j_{\text{bad}})})\}$ has probability at most δ. Only the first unsafe candidate's CS matters, so no union bound over candidates or monotonicity of their risks is required. The CS for $\lambda^{(j_{\text{bad}})}$ is valid because $Z_i(\lambda^{(j_{\text{bad}})})$ has conditional mean $R(\lambda^{(j_{\text{bad}})})$ for every i regardless of which candidate was deployed at i (the audit coin reveals every candidate's loss). □

Corollary 1 (Betting disagreement alarm)**.** *For fixed positive start times $s_1 < s_2 < \dots$ and fixed weights $w_u \ge 0$ with $\sum_u w_u \le 1$, fix $\theta \in (0,1)$ and $\kappa \in (0,1)$. Let $K_t^{(u)}(\theta) = \prod_{s_u \le i \le t} \big(1 + \eta_i(\theta)(Z_i - \theta)\big)$ (equal to 1 for $t < s_u$) with $\mathcal{F}_{i-1}$-measurable $\eta_i(\theta) \in [0, \kappa/(\theta - \ell_i)]$, where $\ell_i < \theta$ is an $\mathcal{F}_{i-1}$-measurable lower bound for Z_i, and let $M_t(\theta) = \sum_u w_u K_t^{(u)}(\theta)$. (i)* Fixed-level alarm. *For a fixed θ, under $H_0(\theta)$: $R_i(\lambda_i) \le \theta \ \forall i$, $M_t(\theta)$ is a nonnegative supermartingale with $M_0 \le 1$, so raising an alarm at the first t with $M_t(\theta) \ge 1/\delta$ has false-alarm probability at most δ over the whole deployment, under arbitrary drift. (ii)* Simultaneous lower bound. *For a fixed grid $0 < \theta_1 < \cdots < \theta_{G_\theta} < 1$, with probability $\ge 1 - \delta$ simultaneously for all t, every grid point θ_g with $M_t(\theta_g) \ge G_\theta/\delta$ satisfies $\theta_g < \max_{i \le t} R_i(\lambda_i)$; the largest rejected grid point is a lower bound on the running maximum risk, uniformly in t (the grid correction is needed because the running maximum is random). Restarts limit dilution by pre-change observations: a component started after a change accumulates only subsequent evidence, with weight penalty $\log(1/w_u)$. We use (i) at $\theta = \alpha$ for alarms and (ii) to plot a lower bound on the running maximum.*

Proof of Corollary 1. Since $Z_i \ge \ell_i$ and $\eta_i(\theta) \le \kappa/(\theta - \ell_i)$, each factor satisfies $1 + \eta_i(\theta)(Z_i - \theta) \ge 1 - \kappa > 0$, so $K_t(\theta)$ is nonnegative. Its conditional increment is $\mathbb{E}_{i-1}[1 + \eta_i(\theta)(Z_i - \theta)] = 1 + \eta_i(\theta)(R_i(\lambda_i) - \theta) \le 1$ whenever $R_i(\lambda_i) \le \theta$, using $\mathbb{E}_{i-1} Z_i = R_i(\lambda_i)$ and predictability of η_i. Under $H_0(\theta)$ it is therefore a supermartingale with $K_0 = 1$; each $K_t^{(u)}(\theta)$ is a nonnegative supermartingale from s_u on and constant before, so $M_t(\theta)$ is a nonnegative supermartingale with $M_0 = \sum_u w_u \le 1$, and Ville's inequality gives $\mathbb{P}[\exists t : M_t(\theta) \ge 1/\delta] \le \delta$ for each fixed θ; this is (i). For (ii), fix a grid point θ_g and let $\sigma_g = \min\{i : R_i(\lambda_i) > \theta_g\}$. Since $R_i(\lambda_i)$ is $\mathcal{F}_{i-1}$-measurable, $\sigma_g - 1$ is a stopping time, and the capital stopped at $\sigma_g - 1$, $M_{t \wedge (\sigma_g - 1)}(\theta_g)$, is a nonnegative supermartingale under *every* data-generating process (before σ_g each factor has conditional expectation at most one; after it the process is constant). Ville's inequality at level δ/G_θ and a union bound over the grid give an event of probability $\ge 1 - \delta$ on which $M_{t \wedge (\sigma_g - 1)}(\theta_g) < G_\theta/\delta$ for all t and g. On this event, if $\max_{i \le t} R_i(\lambda_i) \le \theta_g$ then $t < \sigma_g$, the stopped and the actual capital agree through t, and so $M_t(\theta_g) < G_\theta/\delta$; contrapositively $M_t(\theta_g) \ge G_\theta/\delta$ implies $\max_{i \le t} R_i(\lambda_i) > \theta_g$, a violation *by time* t, which is what (ii) claims. The union over the grid cannot be dropped: the null one would like to test, $\theta = \max_{i \le t} R_i$, is random. The implementation plots the grid point just below the smallest accepted null (at most the largest rejected point, and 0 when nothing is rejected). The same argument with $\eta_i^{\text{up}}(\theta) \le \kappa/(h_i - \theta)$ and factors $1 + \eta_i^{\text{up}}(\theta)(\theta - Z_i)$, h_i the step's upper range, gives the upper bound of the stationary betting CS with per-step clipping. □

Proof of Proposition 2. Under the stated pair, the pre-audit record and conditional audit-coin laws coincide, and the suffix is ancillary given the loss. The conditional divergence of one observation under P_1 relative to P_0 is therefore $p_i \mathrm{kl}(r, \alpha) \le p\mathrm{kl}(r, \alpha)$. For the bounded stopping time $\tau \wedge n$, the chain rule bounds the relative entropy of the stopped observations by $p\mathrm{kl}(r, \alpha)\mathbb{E}_1[\tau \wedge n]$. The event $\{\tau \le n\}$ is measurable from these observations, so data processing gives

$$p\mathrm{kl}(r, \alpha)\mathbb{E}_1[\tau \wedge n] \ \ge \ \mathrm{kl}\big(P_1(\tau \le n), P_0(\tau \le n)\big).$$

If $P_1(\tau \le n) \ge 1 - \beta > \delta$, monotonicity of binary relative entropy lower-bounds the right-hand side by $\mathrm{kl}(1 - \beta, \delta)$. Using $\mathbb{E}_1[\tau \wedge n] \le n$ proves the finite-horizon statement. Letting $n \to \infty$, monotone convergence on the left and lower semicontinuity of relative entropy on the right give a lower bound of $\mathrm{kl}(P_1(\tau < \infty), P_0(\tau < \infty))$. Since $P_1(\tau < \infty) \ge 1 - \beta > \delta \ge P_0(\tau < \infty)$, monotonicity of binary relative entropy yields $\mathrm{kl}(1 - \beta, \delta)$ and the claim. This applies the sequential information argument (Wald, 1945) to observations selected by audits. □

Proof of Theorem 1. (i) On a selected subject, $\mathbb{E}[\mathrm{LR}_g(L_i) \mid \mathcal{F}_{i-1}, B_i = 1] = R_i\, r_g/\alpha + (1 - R_i)(1 - r_g)/(1 - \alpha)$ is affine in R_i, equals 1 at $R_i = \alpha$ and $(1 - r_g)/(1 - \alpha) < 1$ at $R_i = 0$, hence is at most one whenever $R_i \le \alpha$; on an unselected subject the factor is one. Since B_i is drawn with an $\mathcal{F}_{i-1}$-measurable rate and L_i is $\mathrm{Bern}(R_i)$ given the past, each component's capital is

a nonnegative supermartingale from its start and constant before, the grid average and the weighted sum preserve this, and $M_0^{\mathrm{LR}} \le \sum_u w_u \le 1$; Ville's inequality gives (i). (ii) Fix g and write $W_t = w' G^{-1} \prod_{s' \le i \le t:\, B_i = 1} \mathrm{LR}_g(L_i)$, a single nonnegative term in M_t^{LR}. Set $X_i = B_i \log \mathrm{LR}_g(L_i)$, $b = \log(G/(\delta w'))$, $a = \log(r_g/\alpha)$ and $d = pI_g$. For $i \ge s'$ on $\{\tau \ge i\}$, independence of the coin gives

$$\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = \bar{p}_i \left[R_i \log \frac{r_g}{\alpha} + (1 - R_i) \log \frac{1 - r_g}{1 - \alpha} \right] \ge pI_g = d > 0.$$

The bracket increases in R_i, since $r_g > \alpha$, and equals I_g at $R_i = r$. Let $D = (\tau - s' + 1)^+$ and $S_n = \sum_{i=s'}^{s'+n-1} X_i$, with $S_0 = 0$. For each finite n, conditioning each stopped increment on the past yields $\mathbb{E}[S_{D \wedge n}] \ge d\, \mathbb{E}[D \wedge n]$; the indicator that an increment is included is $\mathcal{F}_{i-1}$-measurable and implies $\tau \ge i$. Before an alarm, $W_t \le M_t^{\mathrm{LR}} < 1/\delta$, hence the corresponding sum is below b. At the alarm the final increment is at most a, so pathwise $S_{D \wedge n} < b + a$ (also when $D = 0$). It follows that $\mathbb{E}[D \wedge n] \le (b + a)/d$. Monotone convergence gives $\mathbb{E}[D] \le (b + a)/d$, and $(\tau - s')^+ \le D$ proves the claim. No condition is imposed on risk after the alarm, so the result allows immediate fallback to $\lambda^{(0)}$ on the next subject. □

Fixed non-uniform alternative weights. Theorem 1 also holds with the grid average replaced by $\sum_{g=1}^{G} v_g \prod_{s_u \le i \le t : B_i = 1} \mathrm{LR}_g(L_i)$, where $v_g > 0$ are fixed before monitoring and $\sum_g v_g = 1$. The delay numerator becomes $\log(1/(\delta w' v_g)) + \log(r_g/\alpha)$. Indeed, a fixed convex combination preserves the supermartingale property; in part (ii), replace w'/G by $w' v_g$ and b by $\log(1/(\delta w' v_g))$. Every other step is unchanged. The original experiments use $v_g = 1/G$; E16 compares two fixed extensions without changing their deployment-wide false-alarm guarantee.

Proof of Proposition 3. $\bar{p}_i$ is $\mathcal{F}_{i-1}$-measurable and lies in $[p_{\mathrm{floor}}, p_{\max}]$, so the audit-rate and predictability requirements of Theorems 2 and 3 and Corollary 1 are preserved with $p_{\min} = p_{\mathrm{floor}}$, with all other assumptions retained. The selection conditions of Theorem 1 are also preserved; its delay bound may use $p = p_{\mathrm{floor}}$ when its remaining conditions hold. The per-step range $[1 - 1/\bar{p}_i, 1/\bar{p}_i]$ contains every increment that can occur at step i and is $\mathcal{F}_{i-1}$-measurable, which is all the betting arguments need. The coin-selection identity follows by taking expectations of $\mathbb{E}[B_i \mid \mathcal{F}_{i-1}] = \bar{p}_i$. Conditional independence gives $\mathbb{E}[B_i \mid \mathcal{F}_{i-1}, S_i, C_i] = \bar{p}_i$; multiplying by S_i or $S_i C_i$ and taking expectations yields the paid-audit and cost identities. □

Task-error relation. For label outputs with exact-disagreement loss, as in our LLM and classifier experiments, any true label z_i satisfies

$$\mathbf{1}\{y_{i,d_i(\lambda_i)} \ne z_i\} - \mathbf{1}\{y_{i,R_{\max}} \ne z_i\} \le \mathbf{1}\{y_{i,d_i(\lambda_i)} \ne y_{i,R_{\max}}\}.$$

Taking conditional expectations given $\mathcal{F}_{i-1}$ shows that the proposed output's task-error rate exceeds the full-budget output's by at most $R_i(\lambda_i)$, without requiring the reference to be correct. A cumulative disagreement bound therefore bounds cumulative average excess task error, not absolute task error. For MRI, the certified quantity is the tolerance-exceedance event in Eq. (1).

We call a change a *reference failure* when it makes the full-budget output worse on average than the output proposed by the deployed stopping rule, under an external task-quality measure. Equation (1) measures their discrepancy but cannot identify which output is better. Reference quality must be assessed separately.

Allocation details. We use current log capital, not its running maximum, to determine the normalised evidence. The two-threshold policy switches from p_{floor} to $p_{\max}$ at evidence $1/3$, returns at zero, and retains its previous state between thresholds. Predictability preserves validity but does not equalise costs: the tested null streams have mean paid audit rate near 0.05 at risk 0.02, rising to 0.062 at risk α under smooth escalation (App. G).

Complete deployment procedure. Algorithm 2 expands Algorithm 1, including the optional stationary selector. Confidence levels for cumulative bounds and certification are separate from the alarm-family budget δ.

Algorithm 2 Complete deployment procedure with optional stationary updates

Input: disagreement target α, alarm error budget δ, confidence levels for cumulative-risk bounds and optional rule certification, initial rule λ_1, audit floor and ceiling, and an optional fixed, replayable candidate grid Λ.
Initialise both alarms at level $\delta/2$, initialise the confidence sequences at their specified levels, and set the fallback flag to false.
for subject $i = 1, 2, \ldots$ **do**
  Using $\mathcal{F}_{i-1}$, take the previously selected rule λ_i and choose the scheduled rate $\bar{p}_i$, betting parameters and ranges, and confidence-sequence centres. Draw $B_i \sim \mathrm{Bernoulli}(\bar{p}_i)$ independently of the current subject conditional on $\mathcal{F}_{i-1}$.
  Run the stopping rule without using B_i to determine its proposed stage or output. Record the proposed stage $d_i = d_i(\lambda_i)$ and output.
  Using the observed prefix and past information, compute and record loss predictions for the deployed rule and, when stationary selection is enabled, the candidate rules.
  if $d_i < R_{\max}$ **then**
    Set $A_i = B_i$ and $p_i = \bar{p}_i$; complete the remaining stages when $A_i = 1$.
  else
    Set $A_i = p_i = 1$.
  end if
  Deliver the full-budget output when $A_i = 1$; otherwise deliver the proposed early output. Account for all acquisition cost.
  Using the recorded predictions, form the AIPW increments in Eq. (3). Update the cumulative-risk bound and betting alarm. If stationary selection is enabled, also update the confidence sequences of replayable candidates using $Z_i(\lambda')$.
  Feed the deployed rule's loss to the LR alarm only when $B_i = 1$, including zero losses from subjects that complete naturally.
  Use revealed losses to update the predictor for future subjects.
  If either alarm reaches $2/\delta$, set the fallback flag to true. Once true, the flag remains true.
  If fallback is latched, set $\lambda_{i+1} = \lambda^{(0)}$. Otherwise, if stationary selection is enabled, choose λ_{i+1} by Theorem 3; else retain $\lambda_{i+1} = \lambda_i$.
end for

The default loss predictor uses the stopping score, clipped to $[0, 1]$; for the classifier, this is one minus the largest softmax probability. We also evaluate online isotonic calibration using only past subjects with complete feedback. Each subject's predictions are recorded before auditing reveals its missing losses, and predictor updates apply only to future subjects. Under the boundedness and measurability conditions of §4.3, validity does not require calibrated predictions; their quality affects efficiency through the AIPW increments.

B VALIDITY CHECKS

Monte-Carlo checks of the guarantees (Table 2; $\delta = 0.1$, $T = 1500$, 600 trials per cell, losses observed only through audits at rate p, coins and losses drawn successively from one seeded generator): the time-uniform miscoverage of both upper bounds stays below δ for every failure rate (betting at most 0.097 at $r = 0.2$, where the bound is tight), the drift-valid bound covers the running average under a step change of the risk from 0.02 to 0.3 while the stationary betting bound miscovers up to 0.093 there, the fixed-level betting and LR disagreement tests' false-alarm rates at their true null stay below δ (at most 0.085 and 0.068), the grid-corrected lower bound never exceeded the running maximum, and the closed-loop selector deployed the unsafe candidate (population risk $0.116 > \alpha = 0.1$) in 2 of 150 trials, within its $\delta = 0.1$ budget. These simulations check coverage and false-alarm control in the tested settings. The LR alarm uses rates fixed before each subject, as required by Theorem 1.

C PRE-SPECIFIED HELD-OUT LLM CHECK

The initial LLM study replays cached GSM8K-test and MATH-500 samples with four constructed shifts. A protocol fixed before generation (v2, 2026-09-11, in the repository) tests transfer to 1000 previously unused GSM8K *train* questions, selected with seed 20260911. We generated 64 samples per question on a different machine, retaining the model, prompt and extractor, with temperature 0.8, top-p 0.95 and seeds $20260911 + 10^5$ qid. Generation uses the Hugging Face backend and bfloat16;

Table 2: Validity checks ($\delta = 0.1$, 600 trials per cell). Stationary losses have risk r and audit rate p; disagreement tests use $\theta = r$. Step drift: risk $0.02 \to 0.3$ at $T/2$, disagreement-test threshold 0.3. Miscoverage and false-alarm entries are fractions of streams.

CS	r	p=0.1	p=0.3
		miscoverage of upper bound / false alarm: betting test / LR-mixture test	
betting	0.01	0.000 / 0.000 / 0.038	0.022 / 0.022 / 0.048
betting	0.05	0.065 / 0.033 / 0.047	0.072 / 0.067 / 0.048
betting	0.2	0.093 / 0.055 / 0.048	0.097 / 0.085 / 0.068
eb	0.01	0.000 / –	0.000 / –
eb	0.05	0.000 / –	0.000 / –
eb	0.2	0.000 / –	0.000 / –
drift $0.02 \to 0.3$, p		miscoverage	false alarm: grid bound / fixed-level
betting	0.1	0.093	0.000 / 0.003
betting	0.3	0.083	0.000 / 0.002
eb	0.1	0.000	–
eb	0.3	0.000	–

the original cache used vLLM. Manifests record question hashes, model snapshots, decoding settings and library versions.

The check includes a top-p 0.9 regeneration (seed 20260912, other settings fixed) and the three constructed bugs applied to the new samples. The threshold calibrated on the first 400 GSM8K-test questions remains frozen. The primary policy uses uniform audits at 0.1 and the LR alarm at $\delta = 0.05$, evaluated by its first-ever alarm on ten streams of 6000 questions with the change at 1500.

The five pre-specified criteria are: clean risk at most 0.05; misrouting detected on every stream with lower-median delay below 200; the parser bug detected on every stream with lower median below 2500; duplication detected on at least half the streams; and, if top-p 0.9 has measured risk at most 0.05, no more than one alarm in ten streams on that condition.

Outcome (Table 4). The frozen rule's risk on the clean held-out pool is 0.003 (it stops after 2.1 samples; plurality-at-64 accuracy is 0.963, versus 0.903 on the original test pool). No alarm fires on the clean stream, and the top-p 0.9 change remains within the disagreement target (measured risk 0.010, plurality accuracy 0.962): no alarm of either kind on any of ten streams at any rate. The label-free monitor is silent too, because the means it watches move by 0.002 and 0.001, within its dead zone. This condition meets the disagreement target without a detectable prefix change; MRI intensity drift meets it despite a visible change. The misrouting bug is caught after 38 questions at $\bar{p}$=0.1 (test pool: 38) and the parser bug after a lower median of 925 (1520; per-stream delays 217–2886), on every stream. The duplication criterion is not met: on this pool the duplicated first sample is almost always right, giving risk $0.018 < \alpha$ and no disagreement alarms at any rate. Four of the five specified criteria hold. The failed criterion required detection of a change whose disagreement remained below the target. The optional regeneration control was run later as part of the real-change study below. Betting alarm and audit policies: Table 5.

Regenerated serving configurations (E10). A second protocol, fixed before generation, evaluates operational changes on the same held-out pool. Each configuration is sampled independently with other settings fixed: no change (regeneration control), float16 instead of bfloat16, temperature 1.0 instead of 0.8, the general-purpose Qwen2.5-1.5B-Instruct instead of the Math model (a routing misconfiguration), an "answer concisely" system prompt, a 512-token cap, and 8-bit weights; a second pre-specified battery (protocol v2, after all configurations in the first battery met the disagreement target) adds a system prompt without the `\boxed` instruction, a 256-token cap and 4-bit weights, chosen for having a mechanism by which the rule could fail.

The protocols label a change "harmful" when the frozen rule's pool disagreement exceeds α, and "benign" otherwise. These labels concern disagreement alone. A change is visible if either monitored mean moves by more than the monitor's dead zone; the specified criteria are the held-out ones (an alarm on at most one stream in ten for a benign change; detection for a harmful one). Table 3 reports nine evaluated configurations, including the regeneration control; eight meet the disagreement target (risks 0.001–0.043 against 0.003 clean; the 4-bit variant was not run, its 8-bit sibling having needed

29 GPU-hours), and no alarm of either kind fires on any of their streams at any rate. Two changes are visible. Replacement with the general model lowers the full-budget plurality accuracy from 0.963 to 0.917 and moves the first-checkpoint score by 0.17: the label-free monitor fires after 93 questions, the audits stay silent, and the frozen rule's risk (0.043) remains within the target.

The 256-token cap is the only configuration exceeding the target: it truncates 57% of the samples before their boxed answer (the 512 cap truncates none that matter), truncated samples vote as noise, the rule's risk rises to 0.133 (above the protocol's second threshold 2α) and the full-budget plurality accuracy falls to 0.620. The monitor fires after 64 questions; the LR alarm after lower medians of 2502, 2120 and 1097 questions at $\bar{p}{=}0.05, 0.1, 0.2$ (bounds 5690, 2845, 1423; detected on 80, 90 and 100% of streams within 4500), the betting alarm after 2049, 1574 and 833 (90, 100, 100%), escalation from floor 0.05 after 784 on every stream; with the guardrail on, the delivered failure rate after the change falls from 0.119 to 0.046 at $\bar{p}{=}0.1$. The monitor fires on both visible changes, while audit alarms trigger only on the one exceeding the disagreement target.

The specified criterion for a change above 2α (detection on every stream at $\bar{p}{=}0.1$, lower median below 2500) holds on the median and fails on one stream in ten, whose LR alarm did not fire within the observation window (the betting alarm caught all ten): eight of nine specified criteria hold.

Three specified predictions were wrong and are reported: temperature 1.0 was predicted harmful and visible (it is neither), the general model harmful and invisible (it is benign and visible), and the prompt without the `\boxed` instruction mildly harmful and visible: the Math model boxes its answer in every one of its $64{,}000$ samples whether asked or not, so the change is a no-op for the extractor (risk 0.006, means moved by less than 0.003). The token cap meets its specified classification: visible, with disagreement risk between 0.05 and 0.2. As with the cascade's pixelation, a change that also degrades the full-budget output is caught as disagreement with it; the certificate says nothing about the full budget's own accuracy.

Table 3: E10 serving configurations: 10 streams of 6000, change at 1500; 64-sample accuracy. Delays: lower medians; "–": no detection within 4500. "Any policy": policy/alarm pairs with detections. Δ: changes in proposed-stage fraction and initial score.

real serving change	rule risk (clean 0.003)	plurality acc. (clean 0.963)	Δ stage / Δ score (dead zone 0.02)	label-free monitor	LR alarm, uniform $\bar{p}$ 0.05	0.1	0.2	alarms under any policy (of 16 policy/alarm pairs)	delivered disagr. ($\bar{p}$=0.1)
regeneration, no change	0.005	0.963	+0.001 / +0.000	–	–	–	–	0	0.004
dtype bfloat16 $\rightarrow$ float16	0.001	0.963	+0.001 / +0.000	–	–	–	–	0	0.001
temperature 0.8 $\rightarrow$ 1.0	0.007	0.966	+0.001 / +0.001	–	–	–	–	0	0.006
replacement with the general 1.5B model	0.043	0.917	+0.062 / +0.171	93	–	–	–	0	0.038
system prompt "answer concisely"	0.009	0.964	+0.000 / -0.002	–	–	–	–	0	0.008
8-bit weights (bitsandbytes)	0.004	0.963	+0.000 / +0.000	–	–	–	–	0	0.004
token cap 1024 $\rightarrow$ 512	0.004	0.963	+0.001 / +0.004	–	–	–	–	0	0.004
token cap 1024 $\rightarrow$ 256	0.133	0.620	+0.106 / +0.208	64	2502 (80%)	2120 (90%)	1097	16	0.119
system prompt without `\boxed`	0.006	0.964	+0.001 / -0.003	–	–	–	–	0	0.005

Table 4: Pre-specified LLM check on 1000 held-out GSM8K-train questions (10 streams of 6000, change at 1500). Top-p 0.9 uses freshly generated samples. Column definitions follow Table 23.

(a) Rule risk and label-free monitoring

Condition	Risk before → after	Samples/question after change	Monitor delay
serving bug: first sample duplicated	0.003→0.018	2.0	–
parser bug: 30% answers blank	0.003→0.147	4.0	91
serving bug: first two responses misrouted	0.003→0.986	2.0	–
real serving change: top-p 0.95 → 0.9, regenerated	0.003→0.010	2.1	–

(b) LR delay (detection fraction) / bound

Condition	$p = 0.05$	$p = 0.1$	$p = 0.2$
serving bug: first sample duplicated	–	–	–
parser bug: 30% answers blank	2599 (90%) / 4623	925 / 2312	527 / 1156
serving bug: first two responses misrouted	70 / 106	38 / 53	18 / 27
real serving change: top-p 0.95 → 0.9, regenerated	–	–	–

(c) Escalated LR and delivered disagreement

Condition	Floor 0.05: delay @ paid rate before/after	Delivered disagreement ($p = 0.1$) frozen → fallback
serving bug: first sample duplicated	– @ 0.05→0.05	0.016 → 0.016
parser bug: 30% answers blank	697 @ 0.05→0.44	0.132 → 0.031
serving bug: first two responses misrouted	49 @ 0.05→0.50	0.884 → 0.010
real serving change: top-p 0.95 → 0.9, regenerated	– @ 0.05→0.05	0.008 → 0.008

Table 5: Held-out check, betting alarm and escalated audits (companion to Table 4).

(a) Risk and betting delay (detection fraction) / heuristic scale

Condition	Risk	$p = 0.05$	$p = 0.1$	$p = 0.2$
serving bug: first sample duplicated	0.018	–	–	–
parser bug: 30% answers blank	0.147	1538 / 2123	881 / 1062	488 / 531
serving bug: first two responses misrouted	0.986	156 / 49	86 / 25	59 / 12
real serving change: top-p 0.95 → 0.9, regenerated	0.010	–	–	–

(b) Escalated betting: delay @ paid rate before/after change

Condition	Floor 0.05	Floor 0.02	Net saving ($p = 0.1$) frozen → fallback
serving bug: first sample duplicated	– @ 0.05→0.05	– @ 0.02→0.02	0.868 → 0.868
parser bug: 30% answers blank	572 @ 0.05→0.44	1585 (90%) @ 0.02→0.32	0.847 → 0.367
serving bug: first two responses misrouted	57 @ 0.05→0.50	106 @ 0.02→0.49	0.868 → 0.225
real serving change: top-p 0.95 → 0.9, regenerated	– @ 0.05→0.05	– @ 0.02→0.02	0.867 → 0.867

D MRI PIPELINE DETAILS

Data and split. M4Raw (Lyu et al., 2023) provides raw multi-coil 0.3 T brain acquisitions with three repetitions of T1 and T2 for 183 subjects. The subject-disjoint split uses 65 subjects to train and early-stop a Noise2Noise DnCNN (Lehtinen et al., 2018; Zhang et al., 2017) ($52/13$), 59 calibration subjects and 59 test subjects. Calibration and test subjects are never used to fit the reconstructor, the score, the candidate grid or the tolerance.

Stopping score. After k of $R_{\max}$ repetitions the counterfactual score draws 32 low-rank Gaussian suffixes matching the residual covariance of the prefix, propagates them through the frozen reconstructor, and reports the fraction of draws in which the q-th largest per-slice normalised root mean squared error (nRMSE) between the current and the counterfactual full-budget image exceeds the tolerance ($q = \lfloor 0.1S \rfloor + 1$ over S slices). Candidates threshold this score (§3); at $R_{\max}{=}3$ there is one checkpoint ($k{=}2$) and the scheduler delivers the requested stage without overshoot, charging a 2 s decision stall.

Tolerance and failure. The tolerance is the 0.75 quantile of change-to-full discrepancies on the 40-subject development pool (0.069 T1, 0.068 T2 nRMSE); a subject fails if either contrast's delivered image exceeds it against the full-budget image of the same reconstructor. The cached raw discrepancies allow any other tolerance to be applied as a threshold; the stricter 0.5 quantile (0.047 T1, 0.052 T2) is used in E2 and E3, where a non-trivial candidate ladder is needed.

Acquisition anomalies. Corruptions act on the repetition stack before reconstruction and scoring, on the same subjects: repetition-order permutation; a progressive multiplicative intensity ramp of up to $+20\%$ or $+40\%$ across repetitions; a rigid in-plane shift of one repetition by up to 4 or 8 pixels; an added phase-encode-shifted ghost at 20% or 40% amplitude; and one repetition replaced by a heavy-noise draw at $1\times$ or $2\times$ its own standard deviation (the low signal-to-noise ratio, or low-SNR, dropout of the last repetition). The deployed rule, score, reconstructor and tolerance are frozen throughout.

Reference diagnostics. For every anomaly we also reconstruct, with the same frozen reconstructor, the clean full-budget image of the same subject and record the q-th per-slice nRMSE of the delivered early image and of the completed (anomalous) full-budget image to it. On the 236 units, the early image under the low-SNR dropout is unchanged (the corrupted repetition is the unobserved third one), exceeds the tolerance against the clean full-budget image on 2.1% of units (clean: 2.1%) and is closer to it than the completed image on 99–100% of units (medians 0.035 against 0.094 and 0.290 at $1\times$ and $2\times$): the failure that (1) counts is a failure of the reference, not of the delivered image. The intensity drifts scale every repetition progressively, so there too the completed image lands further from the clean image than the early one (0.100 against 0.062 at $+20\%$), a global intensity scale rather than content; motion and ghosting corrupt the second repetition, so the early image is the worse one, as (1) says, and the rule defers these subjects. Table 23 reports failure rates against both references where they are comparable.

Delivery-path bugs and population shift. Two anomalies act on the clean repetitions downstream of the score, so prefix, score and stop decisions are exactly those of the clean deployment and the full-budget image is unchanged: the early-exit branch delivers the reconstruction of the stage-$(k{-}1)$ prefix (an off-by-one in the stage index), or the raw two-repetition average without the reconstructor. Under the frozen rule these fail 0.695 and 1.000 of the stopped subjects against the true full-budget image; the stage-1 image exceeds the tolerance on 57% of units (median 0.072 against 0.035 for the stage-2 image). The population shift switches the pool at the change to the 40 subjects (of 118) with the largest relative prefix noise, the standard deviation of the difference of the two observed repetitions over that of their average, everything else frozen: their mean score doubles (0.06 against 0.03), the rule defers 11% of their contrasts instead of 5%, and its risk on this pool is 0.051.

Saving. Net saving is the time-weighted cohort fraction $1 - \sum_i T_i^{\text{policy}} / \sum_i T_i^{\text{full}}$ with nominal 60 s per repetition; audited subjects are charged the full protocol plus the stall.

E CHANGES CONFINED TO THE EARLY CLASSIFIER (E14)

Purpose and fixed design. This experiment separates early-branch degradation from reference failure. The dated repository protocol was written before generating the changed predictions. All four prescribed configurations are reported, including the AlexNet replacement that meets the disagreement target and the weakly detected 160-pixel change. We reran first-stage inference on all 3925 Imagenette validation images, then evaluated only the 3125 images outside the original calibration set. The original permutation seed is 20260911, and the original 800-image calibration selects threshold 0.3 for $1-$max-softmax. No new threshold, predictor or audit-policy parameter is fitted on these outcomes.

Inference and reference. All images use the original resize to 256 and centre crop to 224; the two smaller inputs then resize that crop bilinearly to 160 or 128. Normalization and the 1000-class output space are unchanged. The first stage uses torchvision 0.20.1 ResNet-18 or AlexNet `IMAGENET1K_V1` weights with PyTorch 2.5.1, evaluated on CPU. The second and third stages reuse the original, unchanged ResNet-50 and ConvNeXt-Base outputs. These are controlled configuration changes, not observed production incidents. The 224-pixel reproduction differs from the original stage-1 cache on $1/3925$ top-1 labels, with maximum softmax difference 0.0024; its deployment rule risk remains 0.02272. Stage costs count convolution and linear-layer multiply-adds in billions. Data, weight and cache hashes, inference versions and per-image outputs are saved with the report. The four inference runs took 116 CPU wall-clock seconds in total using six threads.

Table 6: E14 configurations on 3125 deployment images. Stage cost is billions of convolution/linear multiply-adds; risks, errors and defer rates are cohort fractions. The reference is unchanged.

Early classifier	Stage cost	Rule risk	Early error	Ref. error	Defer
ResNet-18, 224 (control)	1.814	0.023	0.116	0.111	0.354
ResNet-18, 160	0.926	0.077	0.157	0.111	0.310
ResNet-18, 128	0.593	0.156	0.219	0.111	0.292
AlexNet, 224	0.714	0.020	0.110	0.111	0.476

Paired policy comparison. Each condition uses 30 streams with replacement, seeds 2000–2029, length 3000 and change at 1500. All ten policies share subject orders and domain-separated audit random numbers. Uniform audits use $p = 0.05$; escalation uses floor 0.05, ceiling 0.5 and exponent 0.5. A single alarm uses $\delta = 0.05$; the two-alarm policy splits that budget equally. In the overview figures, "Adaptive audits" denotes alarm-driven escalation with both alarms; all policy variants remain in Table 10. The IPW baseline sets $f = 0$ and uses predictable range $[0, 1/\bar{p}_i]$; AIPW uses the raw score and predictable range $[1 - 1/\bar{p}_i, 1/\bar{p}_i]$. Natural deferrals have loss zero and enter the LR stream only when selected by the pre-subject audit coin, as in the other experiments. A fallback starts on the subject after its alarm and is latched. All costs include natural deferrals, audits and fallbacks. First-stage costs reflect the actual architecture and resolution; later-stage costs are 4.1 and 15.4 GFLOPs. Cost fractions divide summed policy cost by summed full-cascade cost on the same stream, accounting for the configuration change. Table 10 also reports absolute GFLOPs so that these denominators remain explicit.

The fast evaluator retains only the tested risk-grid point, using the same alarm implementations as the full simulator. Equivalence checks compare audit coins, alarm times, delivered losses and costs for nine policy configurations; a separate check compares the reduced and full change monitors. The original 1001-point grid is retained for burn-in confidence bounds. No alarm or audit policy triggers a false alarm before the change on these streams. On the reproduction control, audit policies add 0.030–0.031 of the full budget and never fall back.

Results and uncertainty. Table 10 gives whole-stream means; Figure 1 shows means and interquartile ranges. At 128 pixels, alarm-driven escalation with both alarms lowers disagreement from 0.0926 to 0.0458 and task error from 0.1678 to 0.1364, at normalized cost 0.5698. Uniform LR reaches 0.0734 disagreement and 0.1550 task error at cost 0.4539. These gains accompany higher acquisition cost in this comparison. At 160 pixels, risk is 0.0771 and escalation detects only $5/30$ streams; the 1500-subject window is inadequate for reliable detection. The two-statistic monitor

misses both resolution changes but falls back on $8/30$ streams under the AlexNet replacement, whose disagreement stays below target. The known configuration change could itself trigger immediate fallback; these results concern detection from the tested output statistics.

Table 9 reports paired mean differences with percentile bootstrap intervals from 10000 resamples (seed 20260922). These are descriptive intervals conditional on the image cohort, not additional independent-image evidence or simultaneous intervals over all comparisons. At uniform rate, neither prediction versus IPW nor combining alarms versus LR resolves an improvement. Under escalation, adding AIPW evidence to LR reduces restricted mean delay by 328 subjects while spending an additional 0.078 of the full budget. This experiment measures a cost–quality trade-off; E15 calibrates these policies to normal-operation acquisition cost on separate images (App. F). PPRM and Active Statistical Inference motivate the risk-estimation components, but these component baselines do not reproduce their original batching, query optimization or monitoring targets.

Component diagnostics. The same 30 streams compare LR, inverse-propensity (IPW) betting and prediction-powered (AIPW) betting. In Table 7, "both" combines LR and AIPW; "Escalate" uses alarm evidence to raise the audit rate, while "Joint" also uses the change monitor's evidence. Both escalation variants wait for an audit alarm before fallback.

At uniform rate 0.05, neither AIPW versus IPW nor the two-alarm combination versus LR shows a resolved improvement in whole-stream delivered disagreement; paired 95% bootstrap intervals include zero. Under escalation, the combined LR–AIPW policy reduces restricted mean delay by 328 subjects relative to LR-only escalation, with a descriptive 95% interval of $[208, 458]$. Whole-stream delivered disagreement falls by 0.018 and task error by 0.012, while cost rises by 0.078 (App. E). This comparison changes both the evidence used for allocation and the alarm combination.

A separate diagnostic on the deployment image pools, at uniform audit rate 0.05, finds that the raw predictor reduces AIPW estimator variance relative to IPW by 18–19% under the resolution changes but increases it by 8% under model replacement (App. F). Prediction affects estimator efficiency, but variance reduction alone does not guarantee improved alarm performance.

Table 7: Early-classifier changes at 128 pixels: post-change means over 30 paired streams. Cost is relative to post-change full completion. Restricted mean delay (RMD) is truncated at 1500 subjects; "Detected" counts alarms within that window; dashes mark policies without a detector. Audit rate or floor 0.05, ceiling 0.5.

Policy	Cost	Disagreement	Task error	RMD	Detected
No action	0.312	0.161	0.221	–	–
Always complete	1.000	0.000	0.111	–	–
Change fallback	0.312	0.161	0.221	1500	0/30
Uniform LR	0.468	0.124	0.196	1223	16/30
Uniform IPW	0.456	0.127	0.198	1249	13/30
Uniform AIPW	0.449	0.128	0.198	1265	18/30
Uniform both	0.460	0.126	0.197	1240	15/30
Escalate LR	0.545	0.106	0.183	1065	22/30
Escalate both	0.706	0.069	0.159	737	27/30
Joint both	0.709	0.069	0.158	732	27/30
Window KS (follow-up)	0.789	0.049	0.144	461	30/30

Post-primary distributional-monitor check. After inspecting the primary results, we fixed a stronger label-free control to test whether the missed changes depended on monitoring means. The first 800 deployment subjects form the reference, independently of offline calibration. At subjects $1300, 1350, \ldots, 3000$, two-sided two-sample KS tests (Hodges, 1958) compare the first score and proposed stage with their most recent 500 observations. Reference and test samples are disjoint. Each of the 70 tests uses level $0.05/70$, giving a finite-horizon false-alarm bound when subjects are i.i.d. under a common distribution; overlap between later test windows does not affect the union bound. We use SciPy's exact continuous-null KS tail, which is conservative with ties: applying a common quantile map to two uniform samples restricts the CDF comparison to a subset of the

uniform thresholds. This control tests distributional stability, whereas the audit alarms test risk. Fallback follows the same timing as the primary change-monitor policy.

Table 8 reports all four conditions. KS detects $24/30$ and $30/30$ streams under the two resolution changes, but also triggers on all AlexNet streams despite their low disagreement. On the severe change it reaches lower disagreement than alarm-driven audits at slightly higher cost ($0.0366/0.5943$ versus $0.0458/0.5698$, disagreement/cost). On the AlexNet replacement it costs 0.6423 versus 0.4830, with task errors 0.1124 and 0.1116. It never triggers on the reproduction control or before a change. This follow-up revises the interpretation: the original mean monitor's failure is baseline-dependent. The evidence supports a trade-off between immediate response to change and paying to confirm excess disagreement. It does not establish a general detection advantage over label-free monitoring. The follow-up protocol and results are recorded separately from the primary analysis.

Table 8: Post-primary E14 control: windowed KS tests with direct fallback, 30 paired streams per condition. Column definitions follow Table 10.

Early classifier	Cost	Disagreement	Task error	RMD	Alarm
ResNet-18, 224	0.411	0.024	0.115	1500	0/30
ResNet-18, 160	0.517	0.034	0.126	852	24/30
ResNet-18, 128	0.594	0.037	0.129	461	30/30
AlexNet, 224	0.642	0.014	0.112	343	30/30

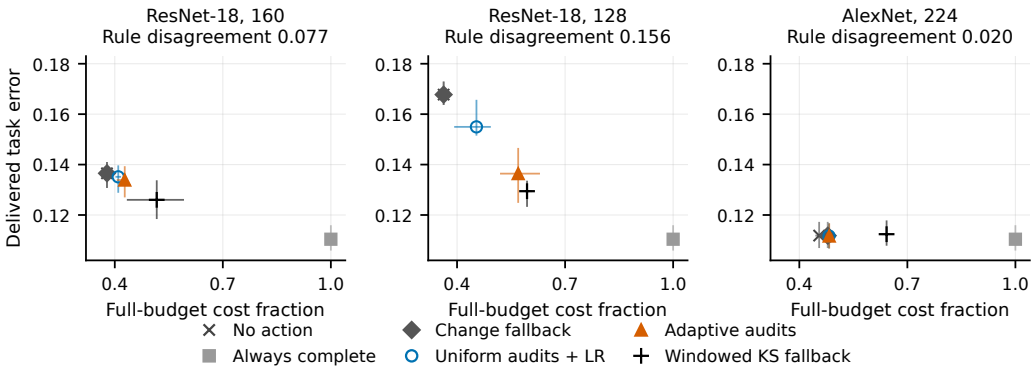

Figure 2: Delivered task error for the six policies in Figure 1. Means and interquartile ranges over the same 30 paired streams; the reference is unchanged.

Table 9: E14, 128 pixels: paired differences (first policy minus second) with 95% bootstrap intervals over 30 streams. Negative values mean less cost, error or delay. U: uniform audits; E: alarm-driven escalation; Monitor: change-triggered fallback.

Comparison	Δ cost	Δ disagr.	Δ task error	Δ RMD
U-AIPW − U-IPW	-0.003 $[-0.024, 0.012]$	0.001 $[-0.003, 0.006]$	0.000 $[-0.002, 0.004]$	16 $[-59, 115]$
U-both − U-LR	-0.004 $[-0.010, 0.003]$	0.001 $[-0.001, 0.003]$	0.001 $[-0.001, 0.002]$	17 $[-16, 49]$
E-LR − U-LR	0.037 $[0.005, 0.070]$	-0.009 $[-0.018, -0.001]$	-0.006 $[-0.012, -0.001]$	-158 $[-314, -3]$
E-both − E-LR	0.078 $[0.053, 0.106]$	-0.018 $[-0.025, -0.012]$	-0.012 $[-0.017, -0.008]$	-328 $[-458, -208]$
Joint − Monitor	0.208 $[0.180, 0.234]$	-0.047 $[-0.054, -0.040]$	-0.031 $[-0.036, -0.027]$	-768 $[-886, -640]$
E-both − Monitor	0.207 $[0.178, 0.233]$	-0.047 $[-0.054, -0.039]$	-0.031 $[-0.036, -0.026]$	-763 $[-883, -635]$

Table 10: E14: means over 30 paired streams. Cost: fraction of full completion; GFLOPs: per subject. Errors concern delivered outputs. RMD truncates delays at 1500; "Alarm" counts post-change fallbacks. KS results: Table 8.

Policy	Cost	Disagr.	Task error	GFLOPs	RMD	Alarm
ResNet-18, 224 (control)						
No action	0.411	0.024	0.115	8.77	1500	0/30
Always complete	1.000	0.000	0.110	21.31	1500	0/30
Change fallback	0.411	0.024	0.115	8.77	1500	0/30
Uniform LR	0.441	0.022	0.115	9.41	1500	0/30
Uniform IPW	0.441	0.022	0.115	9.41	1500	0/30
Uniform AIPW	0.441	0.022	0.115	9.41	1500	0/30
Uniform both	0.441	0.022	0.115	9.41	1500	0/30
Escalate LR	0.442	0.022	0.115	9.41	1500	0/30
Escalate both	0.442	0.022	0.115	9.42	1500	0/30
Joint both	0.442	0.022	0.115	9.42	1500	0/30
ResNet-18, 160						
No action	0.378	0.052	0.137	7.89	1500	0/30
Always complete	1.000	0.000	0.110	20.86	1500	0/30
Change fallback	0.378	0.052	0.137	7.89	1500	0/30
Uniform LR	0.410	0.049	0.135	8.55	1500	0/30
Uniform IPW	0.410	0.049	0.135	8.55	1500	0/30
Uniform AIPW	0.411	0.049	0.135	8.58	1493	1/30
Uniform both	0.410	0.049	0.135	8.55	1499	1/30
Escalate LR	0.410	0.049	0.135	8.55	1500	0/30
Escalate both	0.428	0.047	0.134	8.92	1452	5/30
Joint both	0.428	0.047	0.134	8.93	1452	5/30
ResNet-18, 128						
No action	0.363	0.093	0.168	7.51	1500	0/30
Always complete	1.000	0.000	0.110	20.70	1500	0/30
Change fallback	0.363	0.093	0.168	7.51	1500	0/30
Uniform LR	0.454	0.073	0.155	9.39	1223	16/30
Uniform IPW	0.448	0.075	0.156	9.27	1249	13/30
Uniform AIPW	0.445	0.076	0.156	9.20	1265	18/30
Uniform both	0.450	0.074	0.156	9.31	1240	15/30
Escalate LR	0.491	0.064	0.149	10.17	1065	22/30
Escalate both	0.570	0.046	0.136	11.79	737	27/30
Joint both	0.571	0.046	0.136	11.82	732	27/30
AlexNet, 224						
No action	0.455	0.022	0.112	9.45	1500	0/30
Always complete	1.000	0.000	0.110	20.76	1500	0/30
Change fallback	0.479	0.021	0.112	9.94	1352	8/30
Uniform LR	0.482	0.021	0.112	10.01	1500	0/30
Uniform IPW	0.482	0.021	0.112	10.01	1500	0/30
Uniform AIPW	0.482	0.021	0.112	10.01	1500	0/30
Uniform both	0.482	0.021	0.112	10.01	1500	0/30
Escalate LR	0.483	0.021	0.112	10.02	1500	0/30
Escalate both	0.483	0.021	0.112	10.03	1500	0/30
Joint both	0.502	0.020	0.112	10.42	1500	0/30

F FOLLOW-UP AUDIT-BUDGET AND GRID CHECKS

These follow-ups address comparisons suggested by the primary results. Their protocol, seeds, grids, calibration targets and primary contrasts were recorded in a protocol fixed before execution. The original E1–E14 results and LR defaults are retained.

F.1 ACQUISITION BUDGETS ON CLASSIFIER OUTPUTS (E15)

E15 uses the original E14 predictions and stopping threshold. The original permutation (seed 20260911) assigns 800 images to calibration and the remaining 3125 to evaluation. For each policy, 48 unchanged streams of length 3000 (seeds 41000–41047) calibrate the audit-rate or floor parameter using only the first pool. Evaluation uses 100 streams (seeds 42000–42099), with a change at 1500, from the second pool. Subject orders and audit uniforms are shared across policies. Resampling is with replacement; intervals describe stream variation conditional on the fixed pools.

The budget is additional acquisition cost over the unaudited frozen rule, divided by full-cascade cost on the same stream. It includes any false-alarm fallback during calibration. Targets are 0.015, 0.03 and 0.06, with 0.03 designated primary. Each calibrated parameter lies in $[0.001, 0.25]$; the search tests an initial proportional-cost guess, endpoints if needed, and up to eight bisection steps, retaining the closest tested value. The absolute tolerance is 0.0005. Finite-stream costs need not be monotone in this parameter because an alarm changes all subsequent costs. The lowest-target IPW run misses tolerance: its closest tested cost is 0.02066, so it is not treated as a matched-budget point. All other calibration targets meet tolerance (Table 11).

The eight families combine uniform, smooth or two-threshold audit rates with LR, IPW, AIPW or both alarms. Smooth escalation uses exponent 0.5; the two-threshold controller switches to 0.5 at evidence $1/3$ and returns to its floor at zero. A single alarm receives $\delta = 0.05$; both alarms split this level. The ceiling, thresholds and raw predictor are fixed across budgets. No predictor is fitted to changed outcomes. No-action, always-complete and windowed KS policies are reference comparisons; KS tests a distributional null and is not audit-budget matched.

Table 11: E15 budget calibration. Cal. alarms count 48 unchanged streams; pre alarms count 100 evaluation prefixes, shared across configurations. "Met" denotes absolute target error at most 0.0005.

Target	Policy	Floor	Cal. cost	Cal. alarms	Pre alarms	Met
0.015	Uniform LR	0.02464	0.01454	0	1	Yes
0.015	Uniform IPW	0.01309	0.02066	1	0	No
0.015	Uniform AIPW	0.02464	0.01454	0	0	Yes
0.015	Uniform both	0.02464	0.01454	0	0	Yes
0.015	Smooth LR	0.02464	0.01461	0	0	Yes
0.015	Smooth both	0.02464	0.01516	0	0	Yes
0.015	Two-threshold LR	0.02464	0.01456	0	0	Yes
0.015	Two-threshold both	0.02464	0.01470	0	0	Yes
0.030	Uniform LR	0.04927	0.03023	0	1	Yes
0.030	Uniform IPW	0.04927	0.03023	0	0	Yes
0.030	Uniform AIPW	0.04927	0.03023	0	0	Yes
0.030	Uniform both	0.04927	0.03023	0	0	Yes
0.030	Smooth LR	0.04927	0.03033	0	0	Yes
0.030	Smooth both	0.04776	0.02981	0	0	Yes
0.030	Two-threshold LR	0.04927	0.03025	0	0	Yes
0.030	Two-threshold both	0.04927	0.03032	0	0	Yes
0.060	Uniform LR	0.09854	0.06015	0	0	Yes
0.060	Uniform IPW	0.09854	0.06015	0	1	Yes
0.060	Uniform AIPW	0.09854	0.06015	0	0	Yes
0.060	Uniform both	0.09854	0.06015	0	0	Yes
0.060	Smooth LR	0.09854	0.06017	0	0	Yes
0.060	Smooth both	0.09702	0.05969	0	0	Yes
0.060	Two-threshold LR	0.09854	0.06015	0	0	Yes
0.060	Two-threshold both	0.09854	0.06015	0	0	Yes

Primary comparisons and uncertainty. At 128 pixels and target 0.03, the three designated comparators are uniform LR, smooth LR and two-threshold both. The two primary outcomes are post-change disagreement and restricted mean delay, giving six contrasts. Table 12 reports smooth both minus each comparator using 10000 paired bootstrap resamples (seed 20260923). Alongside descriptive 95% intervals, percentile intervals at level $1 - 0.05/6$ provide the Bonferroni adjustment for a nominal 95% familywise level. This adjustment does not make bootstrap coverage exact in finite samples. All six adjusted intervals exclude zero. For the uniform-LR delay contrast, one pre-change alarm is excluded from both paired vectors, leaving 99 pairs; the other delay contrasts use 100. Every stream, including pre-change alarms, remains in delivered-outcome and cost summaries.

Smooth both reduces post-change disagreement by 0.0518, 0.0362 and 0.0159, and restricted mean delay by 477, 329 and 149 subjects. Its post-change cost is 0.665 of full acquisition, compared with 0.440, 0.507 and 0.597. Realised pre-change extra costs are 0.0302, 0.0336, 0.0293 and 0.0297, respectively; the uniform-LR mean includes its one false-alarm fallback. This is a calibration-target comparison with separate changed-period expenditure, not an equal-total-cost comparison. Post-change task error is 0.163 for smooth both versus 0.199, 0.188 and 0.174 for these comparators.

Table 12: E15 primary paired differences: smooth both minus comparator. Intervals use 10000 paired bootstrap resamples; the adjusted column accounts for six contrasts. Negative differences favour smooth both.

Comparator	Outcome	Difference	95% interval	Simultaneous interval	n
Uniform LR	Post disagr.	-0.052	[-0.060, -0.044]	[-0.062, -0.042]	100
Uniform LR	RMD	-477	[-551, -406]	[-577, -384]	99
Smooth LR	Post disagr.	-0.036	[-0.043, -0.030]	[-0.045, -0.028]	100
Smooth LR	RMD	-329	[-391, -269]	[-413, -250]	100
Two-threshold both	Post disagr.	-0.016	[-0.020, -0.011]	[-0.022, -0.010]	100
Two-threshold both	RMD	-149	[-195, -104]	[-211, -88]	100

Figure 3 shows all three budgets for the four primary families. Tables 13–16 retain every configuration, budget and policy. Costs and outcomes include natural deferrals, audits and fallback. RMD is truncated at 1500 and excludes pre-change alarms; Table 11 gives the excluded counts, so its denominator is 100 minus that count. "Alarm" counts post-change alarms even when disagreement stays below target; it does not by itself establish excess disagreement. A mean post-change disagreement of d corresponds to $1500d$ delivered failures in that period. Per-stream reports also retain failures before intervention, audit counts and alarm times. All settings are reported; none was selected using changed-period outcomes after budget calibration.

Table 13: E15 reproduction control (ResNet-18, 224 pixels): means over 100 streams. "Pre extra" is added cost before the change; "All" and "Post" specify the measurement period. RMD is truncated at 1500.

Policy	Pre extra	All cost	Post cost	Post dis.	Post error	RMD	Alarm
Target extra cost 0.015							
Uniform LR	0.019	0.431	0.433	0.022	0.115	1500	0/100
Uniform IPW	0.008	0.420	0.420	0.022	0.115	1500	0/100
Uniform AIPW	0.015	0.426	0.427	0.022	0.115	1500	0/100
Uniform both	0.015	0.426	0.427	0.022	0.115	1500	0/100
Smooth LR	0.015	0.426	0.427	0.022	0.115	1500	0/100
Smooth both	0.017	0.428	0.427	0.022	0.115	1500	0/100
Two-threshold LR	0.015	0.426	0.427	0.022	0.115	1500	0/100
Two-threshold both	0.015	0.427	0.427	0.022	0.115	1500	0/100
Target extra cost 0.03							
Uniform LR	0.034	0.446	0.447	0.021	0.115	1500	0/100
Uniform IPW	0.029	0.441	0.441	0.021	0.115	1500	0/100
Uniform AIPW	0.029	0.441	0.441	0.021	0.115	1500	0/100
Uniform both	0.029	0.441	0.441	0.021	0.115	1500	0/100
Smooth LR	0.029	0.441	0.441	0.021	0.115	1500	0/100
Smooth both	0.030	0.441	0.441	0.021	0.115	1500	0/100
Two-threshold LR	0.029	0.441	0.441	0.021	0.115	1500	0/100
Two-threshold both	0.030	0.441	0.441	0.021	0.115	1500	0/100
Target extra cost 0.06							
Uniform LR	0.058	0.470	0.471	0.020	0.115	1500	0/100
Uniform IPW	0.064	0.475	0.476	0.020	0.115	1500	0/100
Uniform AIPW	0.058	0.470	0.471	0.020	0.115	1500	0/100
Uniform both	0.058	0.470	0.471	0.020	0.115	1500	0/100
Smooth LR	0.059	0.470	0.471	0.020	0.115	1500	0/100
Smooth both	0.059	0.470	0.470	0.020	0.115	1500	0/100
Two-threshold LR	0.059	0.470	0.471	0.020	0.115	1500	0/100
Two-threshold both	0.059	0.470	0.471	0.020	0.115	1500	0/100
No action	0.000	0.412	0.413	0.022	0.115	–	0/100
Always complete	0.589	1.000	1.000	0.000	0.112	–	0/100
Window KS	0.000	0.415	0.418	0.022	0.115	1486	1/100

Table 14: E15, ResNet-18 at 160 pixels: means over 100 streams. Costs are relative to full acquisition in the indicated period; post dis. and post error are delivered outcomes after the change.

Policy	Pre extra	All cost	Post cost	Post dis.	Post error	RMD	Alarm
Target extra cost 0.015							
Uniform LR	0.019	0.402	0.373	0.074	0.155	1487	1/100
Uniform IPW	0.008	0.390	0.359	0.075	0.156	1489	1/100
Uniform AIPW	0.015	0.395	0.362	0.075	0.155	1500	0/100
Uniform both	0.015	0.397	0.367	0.074	0.155	1488	1/100
Smooth LR	0.015	0.396	0.364	0.075	0.155	1498	1/100
Smooth both	0.017	0.408	0.388	0.072	0.154	1468	7/100
Two-threshold LR	0.015	0.395	0.362	0.075	0.155	1500	0/100
Two-threshold both	0.015	0.399	0.370	0.074	0.155	1493	2/100
Target extra cost 0.03							
Uniform LR	0.034	0.415	0.384	0.072	0.154	1500	0/100
Uniform IPW	0.029	0.410	0.378	0.073	0.154	1500	0/100
Uniform AIPW	0.029	0.410	0.379	0.073	0.154	1497	1/100
Uniform both	0.029	0.410	0.378	0.073	0.154	1500	0/100
Smooth LR	0.029	0.412	0.383	0.072	0.154	1489	1/100
Smooth both	0.030	0.424	0.406	0.070	0.152	1464	8/100
Two-threshold LR	0.029	0.410	0.379	0.073	0.154	1498	2/100
Two-threshold both	0.030	0.420	0.399	0.071	0.153	1469	7/100
Target extra cost 0.06							
Uniform LR	0.058	0.441	0.411	0.069	0.152	1500	0/100
Uniform IPW	0.064	0.446	0.417	0.068	0.151	1500	0/100
Uniform AIPW	0.058	0.445	0.419	0.068	0.151	1481	4/100
Uniform both	0.058	0.443	0.415	0.069	0.152	1490	2/100
Smooth LR	0.059	0.442	0.413	0.069	0.152	1500	0/100
Smooth both	0.059	0.453	0.436	0.066	0.150	1465	7/100
Two-threshold LR	0.059	0.442	0.414	0.069	0.152	1494	3/100
Two-threshold both	0.059	0.455	0.439	0.066	0.150	1462	8/100
No action	0.000	0.379	0.346	0.077	0.157	–	0/100
Always complete	0.589	1.000	1.000	0.000	0.112	–	0/100
Window KS	0.000	0.519	0.633	0.043	0.136	841	78/100

Table 15: E15, ResNet-18 at 128 pixels: all budgets and policies, means over 100 streams. RMD excludes pre-change alarms listed in Table 11; all streams contribute to costs and outcomes.

Policy	Pre extra	All cost	Post cost	Post dis.	Post error	RMD	Alarm
Target extra cost 0.015							
Uniform LR	0.019	0.399	0.367	0.144	0.210	1440	13/100
Uniform IPW	0.008	0.376	0.330	0.153	0.216	1490	2/100
Uniform AIPW	0.015	0.387	0.345	0.149	0.213	1474	9/100
Uniform both	0.015	0.392	0.356	0.147	0.212	1451	10/100
Smooth LR	0.015	0.433	0.441	0.127	0.198	1271	46/100
Smooth both	0.017	0.498	0.573	0.097	0.177	1008	72/100
Two-threshold LR	0.015	0.422	0.417	0.132	0.202	1320	37/100
Two-threshold both	0.015	0.440	0.456	0.124	0.196	1254	48/100
Target extra cost 0.03							
Uniform LR	0.034	0.442	0.440	0.128	0.199	1305	34/100
Uniform IPW	0.029	0.431	0.422	0.132	0.201	1333	29/100
Uniform AIPW	0.029	0.441	0.441	0.127	0.198	1291	46/100
Uniform both	0.029	0.432	0.424	0.131	0.201	1329	36/100
Smooth LR	0.029	0.472	0.507	0.112	0.188	1154	63/100
Smooth both	0.030	0.550	0.665	0.076	0.163	825	90/100
Two-threshold LR	0.029	0.464	0.489	0.116	0.191	1193	55/100
Two-threshold both	0.030	0.516	0.597	0.092	0.174	974	81/100
Target extra cost 0.06							
Uniform LR	0.058	0.507	0.547	0.103	0.182	1104	73/100
Uniform IPW	0.064	0.508	0.544	0.104	0.183	1121	65/100
Uniform AIPW	0.058	0.536	0.607	0.089	0.173	956	94/100
Uniform both	0.058	0.523	0.579	0.095	0.177	1026	83/100
Smooth LR	0.059	0.541	0.617	0.087	0.171	952	86/100
Smooth both	0.059	0.609	0.757	0.054	0.149	633	99/100
Two-threshold LR	0.059	0.542	0.620	0.086	0.171	941	84/100
Two-threshold both	0.059	0.593	0.724	0.062	0.155	709	99/100
No action	0.000	0.365	0.317	0.156	0.218	–	0/100
Always complete	0.589	1.000	1.000	0.000	0.112	–	0/100
Window KS	0.000	0.591	0.782	0.049	0.145	478	99/100

Table 16: E15, AlexNet replacement: means over 100 streams. This change is within the disagreement target; alarms and fallback need not indicate excess disagreement.

Policy	Pre extra	All cost	Post cost	Post dis.	Post error	RMD	Alarm
Target extra cost 0.015							
Uniform LR	0.019	0.473	0.519	0.019	0.110	1500	0/100
Uniform IPW	0.008	0.462	0.508	0.020	0.110	1500	0/100
Uniform AIPW	0.015	0.469	0.514	0.019	0.110	1500	0/100
Uniform both	0.015	0.469	0.514	0.019	0.110	1500	0/100
Smooth LR	0.015	0.469	0.514	0.019	0.110	1500	0/100
Smooth both	0.017	0.470	0.514	0.019	0.110	1500	0/100
Two-threshold LR	0.015	0.469	0.514	0.019	0.110	1500	0/100
Two-threshold both	0.015	0.469	0.514	0.019	0.110	1500	0/100
Target extra cost 0.03							
Uniform LR	0.034	0.486	0.531	0.019	0.110	1500	0/100
Uniform IPW	0.029	0.482	0.526	0.019	0.110	1500	0/100
Uniform AIPW	0.029	0.482	0.526	0.019	0.110	1500	0/100
Uniform both	0.029	0.482	0.526	0.019	0.110	1500	0/100
Smooth LR	0.029	0.482	0.526	0.019	0.110	1500	0/100
Smooth both	0.030	0.482	0.525	0.019	0.110	1500	0/100
Two-threshold LR	0.029	0.482	0.526	0.019	0.110	1500	0/100
Two-threshold both	0.030	0.482	0.526	0.019	0.110	1500	0/100
Target extra cost 0.06							
Uniform LR	0.058	0.509	0.551	0.018	0.110	1500	0/100
Uniform IPW	0.064	0.514	0.555	0.018	0.110	1500	0/100
Uniform AIPW	0.058	0.509	0.551	0.018	0.110	1500	0/100
Uniform both	0.058	0.509	0.551	0.018	0.110	1500	0/100
Smooth LR	0.059	0.509	0.551	0.018	0.110	1500	0/100
Smooth both	0.059	0.509	0.550	0.018	0.110	1500	0/100
Two-threshold LR	0.059	0.509	0.551	0.018	0.110	1500	0/100
Two-threshold both	0.059	0.509	0.551	0.018	0.110	1500	0/100
No action	0.000	0.455	0.502	0.020	0.110	–	0/100
Always complete	0.589	1.000	1.000	0.000	0.112	–	0/100
Window KS	0.000	0.643	0.887	0.004	0.111	340	100/100

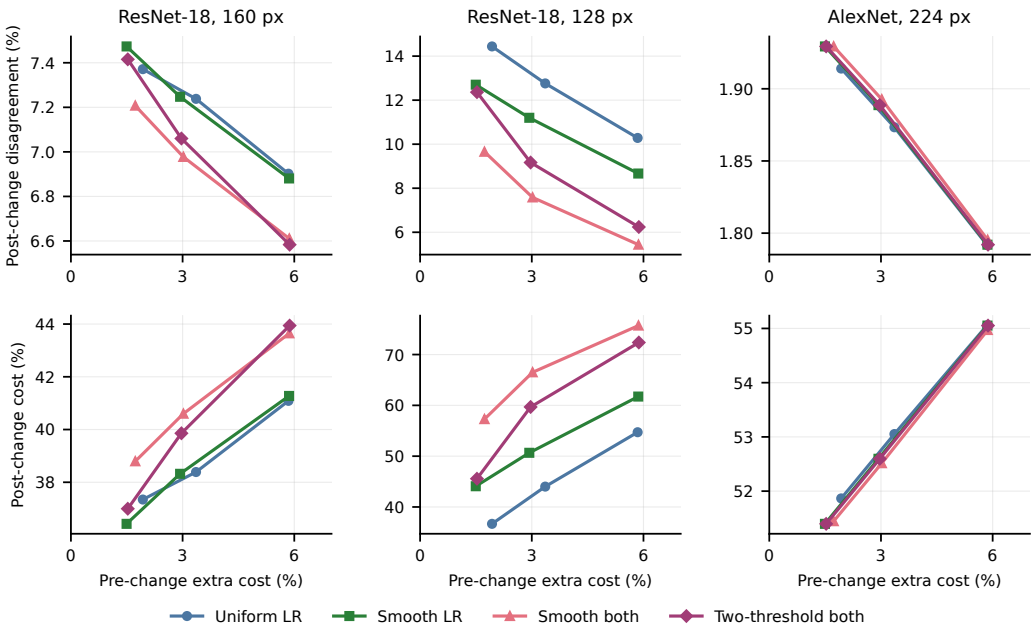

Figure 3: E15: mean post-change outcomes versus realised pre-change extra cost, at all three calibration targets. Each point uses 100 paired streams. Four primary families shown; all policies appear in the tables.

F.2 What does the loss predictor contribute?

On each full deployment pool, we compare the raw score f with the zero predictor, using fixed audit rate 0.05 and propensity one on natural completions. This is a retrospective diagnostic, not predictor selection using changed losses. For a uniformly drawn subject with loss L, the law of total variance gives

$$\mathrm{Var}(Z) = \mathrm{Var}(L) + \mathbb{E}\left[\frac{1-p}{p}(L-f)^2\right], \qquad p = 1 \text{ on natural completions.}$$

Table 17 evaluates this identity on the finite pool. The raw predictor reduces total variance by 18.3% and 19.4% on the two resolution changes, but increases it by 8.4% under AlexNet replacement. On clean data the reduction is only 1.2%. Variance reduction alone does not order the alarms: at rate 0.05 the AIPW increments permit range $[-19, 20]$, versus $[0, 20]$ for IPW. These different ranges constrain betting differently. The E14 and E15 component comparisons therefore assess the resulting alarms, rather than equating improved squared error with improved detection.

Table 17: Predictor diagnostics on each 3125-image deployment pool. MSE is squared error for the rule loss; variances include subject and audit randomness at rate 0.05. Ratio is AIPW variance divided by IPW variance.

Condition	MSE, zero	MSE, raw	Var., IPW	Var., AIPW	Ratio
Original clean	0.0230	0.0227	0.4603	0.4547	0.9878
ResNet-18, 224 (control)	0.0227	0.0226	0.4539	0.4508	0.9933
ResNet-18, 160	0.0771	0.0623	1.5365	1.2557	0.8173
ResNet-18, 128	0.1562	0.1246	3.0988	2.4986	0.8063
AlexNet, 224	0.0198	0.0216	0.3964	0.4296	1.0838

F.3 Alternative-grid sensitivity (E16)

E16 retains the original grid $\{0.1, 0.2, 0.4, 0.8, 0.95\}$ and compares two extensions adding $\{0.055, 0.06, 0.07, 0.08, 0.09\}$. The extended grid assigns equal weight to ten alternatives. The

weighted grid assigns total mass 0.2 to the five new alternatives and 0.8 to the originals. All use $(\alpha, \delta) = (0.05, 0.05)$ and the original restart schedules. The fixed-weight extension of Theorem 1 is proved in App. A; changing these priors does not alter the false-alarm level.

Synthetic streams have constant visible prefixes and Bernoulli losses: risk 0.02 before subject 1500, followed by 6000 subjects at each specified risk. Rates are 0.05 and 0.1, with 100 paired streams (seeds 51000–51099). Secondary changes at 1533 use rate 0.1 and risks 0.077, 0.1 and 0.2. Separate loss and audit random streams are shared across grids. A vectorised evaluator sums the same restarted capital at audit times, with future components at capital one; between audits the total capital is unchanged. Tests compare capitals and first crossings with the streaming implementation for all grids, including zero and full auditing and changes off the restart calendar.

The smallest positive-log-drift threshold decreases from 0.07236 to 0.05246. This narrows the range without positive drift; it does not increase the information per audit. At $r = 0.06$ and $p = 0.1$, the original expected-delay bound is infinite; the extended and weighted bounds are still approximately 131775 and 141010 subjects. Within 6000 subjects, the three grids detect 0, 1 and 0 of 100 streams. At $r = 0.1$ the equal-weight extension improves detection from 75 to 85 streams; at $r = 0.2$ and 0.5 it increases RMD from 806 to 828 and from 148 to 158. The original grid retains more prior mass on severe alternatives. Tables 18–20 report all risks, detection counts, censored delays, failures before intervention and analytic bounds. A censored RMD is not an estimate of the full expectation appearing in Theorem 1, so these columns are not a test of that bound. No pre-change alarm occurs in these synthetic change streams.

Table 18: E16 synthetic changes: RMD truncated at 6000 subjects and detected streams out of 100. All three fixed grids use the same losses and audit coins.

r	p	Change	Original		Extended		Weighted	
			RMD	Det./100	RMD	Det./100	RMD	Det./100
0.055	0.05	1500	5942	1	5983	1	5942	1
0.060	0.05	1500	5925	3	5980	2	5925	3
0.070	0.05	1500	5897	3	5952	2	5896	3
0.077	0.05	1500	5890	4	5931	4	5883	5
0.080	0.05	1500	5825	9	5864	9	5815	10
0.100	0.05	1500	5194	48	5123	56	5120	54
0.200	0.05	1500	1319	100	1390	100	1341	100
0.500	0.05	1500	257	100	276	100	263	100
0.055	0.10	1500	6000	0	6000	0	6000	0
0.060	0.10	1500	6000	0	5997	1	6000	0
0.070	0.10	1500	5830	9	5781	12	5807	11
0.077	0.10	1500	5743	16	5664	19	5702	19
0.080	0.10	1500	5592	20	5446	29	5543	24
0.100	0.10	1500	4057	75	3851	85	3985	77
0.200	0.10	1500	806	100	828	100	809	100
0.500	0.10	1500	148	100	158	100	151	100
0.077	0.10	1533	5787	14	5685	19	5742	19
0.100	0.10	1533	4084	73	3881	84	3982	78
0.200	0.10	1533	793	100	829	100	800	100

Table 19: E16: mean delivered failures after the change and before alarm, or until the horizon if no alarm occurs. Audits deliver zero disagreement.

r	p	Change	Original	Extended	Weighted
0.055	0.05	1500	310.8	312.9	310.8
0.060	0.05	1500	338.5	341.4	338.4
0.070	0.05	1500	393.3	396.7	393.2
0.077	0.05	1500	431.6	434.4	431.1
0.080	0.05	1500	443.9	446.9	443.2
0.100	0.05	1500	495.1	488.1	487.8
0.200	0.05	1500	251.9	265.4	256.3
0.500	0.05	1500	122.7	131.5	125.6
0.055	0.10	1500	297.3	297.3	297.3
0.060	0.10	1500	324.8	324.7	324.8
0.070	0.10	1500	368.9	365.8	367.4
0.077	0.10	1500	399.4	393.7	396.6
0.080	0.10	1500	404.5	393.9	400.9
0.100	0.10	1500	366.6	347.6	359.9
0.200	0.10	1500	145.5	149.4	146.0
0.500	0.10	1500	66.7	71.5	68.2
0.077	0.10	1533	402.5	395.4	399.4
0.100	0.10	1533	369.1	350.1	359.6
0.200	0.10	1533	143.2	149.6	144.4

Table 20: E16: analytic upper bounds on full expected delay, including the wait to the first post-change linear restart. Infinity denotes no positive-drift component.

r	p	Change	Original	Extended	Weighted
0.055	0.05	1500	∞	1017587	1089386
0.060	0.05	1500	∞	263549	282021
0.070	0.05	1500	∞	70265	75132
0.077	0.05	1500	74337	40656	43444
0.080	0.05	1500	45154	33076	35344
0.100	0.05	1500	12483	13154	12699
0.200	0.05	1500	1944	2043	1976
0.500	0.05	1500	353	370	358
0.055	0.10	1500	∞	508794	544693
0.060	0.10	1500	∞	131775	141010
0.070	0.10	1500	∞	35133	37566
0.077	0.10	1500	37169	20328	21722
0.080	0.10	1500	22577	16538	17672
0.100	0.10	1500	6241	6577	6349
0.200	0.10	1500	972	1021	988
0.500	0.10	1500	176	185	179
0.077	0.10	1533	37366	20440	21834
0.100	0.10	1533	6289	6624	6397
0.200	0.10	1533	993	1043	1009

Null controls and classifier checks. At each constant null risk 0.02 or 0.05 and each audit rate, we run 500 streams of length 7500 (seeds 52000–52499). Table 21 reports alarm fractions with exact binomial 95% intervals. Observed fractions range from 0 to 0.024; these finite-horizon results complement the deployment-wide guarantee of Theorem 1.

The secondary classifier check applies all three grids with uniform audit probability 0.1 to the same 100 E15 streams, without recalibrating on their outcomes (Table 22). None detects the 160-pixel condition within 1500 post-change subjects. At 128 pixels, the extended grid slightly increases RMD (1109 versus 1089) and post-change disagreement (0.103 versus 0.101). We retain the original default: the extension improves some moderate-risk cases at the cost of slower detection elsewhere.

Table 21: E16 null controls: alarms within 7500 subjects over 500 streams per cell, with exact binomial 95% intervals.

r	p	Grid	Alarms	Fraction	95% interval
0.02	0.05	Original	2/500	0.004	[0.000, 0.014]
0.02	0.05	Extended	1/500	0.002	[0.000, 0.011]
0.02	0.05	Weighted	2/500	0.004	[0.000, 0.014]
0.05	0.05	Original	9/500	0.018	[0.008, 0.034]
0.05	0.05	Extended	8/500	0.016	[0.007, 0.031]
0.05	0.05	Weighted	8/500	0.016	[0.007, 0.031]
0.02	0.10	Original	0/500	0.000	[0.000, 0.007]
0.02	0.10	Extended	0/500	0.000	[0.000, 0.007]
0.02	0.10	Weighted	0/500	0.000	[0.000, 0.007]
0.05	0.10	Original	12/500	0.024	[0.012, 0.042]
0.05	0.10	Extended	9/500	0.018	[0.008, 0.034]
0.05	0.10	Weighted	11/500	0.022	[0.011, 0.039]

Table 22: E16 classifier checks at audit rate 0.1: means over 100 paired streams. RMD is truncated at 1500; no pre-change alarms occur.

Condition	Grid	Post cost	Post dis.	Post error	RMD	Detected
ResNet-18, 160	Original	0.412	0.069	0.152	1500	0/100
ResNet-18, 160	Extended	0.412	0.069	0.152	1500	0/100
ResNet-18, 160	Weighted	0.412	0.069	0.152	1500	0/100
ResNet-18, 128	Original	0.554	0.101	0.181	1089	73/100
ResNet-18, 128	Extended	0.546	0.103	0.182	1109	73/100
ResNet-18, 128	Weighted	0.551	0.102	0.182	1095	72/100

Table 23: MRI anomalies at subject 1500 of 3000 (10 streams). Delays: lower medians; parentheses: detection fractions below one (monitor column: defer rates). Delivered disagreement compares the frozen rule with fallback using both alarms ($\bar{p} = 0.1$, level $\delta/2$ each).

anomaly	risk after Eq. (1) / clean ref.	label-free monitor (defer rate)	integrity check	LR disagreement alarm $\bar{p}$=0.05	0.1	0.2	delivered disagreement frozen → guardrail
clean control	0.017 / 0.017	– (0.00)	–	–	–	–	0.016 → 0.016
repetition order permuted	0.000 / n/a	– (0.02)	–	–	–	–	0.000 → 0.000
intensity drift +20%	0.025 / n/a	80 (0.16)	–	–	–	–	0.022 → 0.022
intensity drift +40%	0.000 / n/a	36 (1.00)	–	–	–	–	0.000 → 0.000
motion spike 4 / 8 px	0.000 / n/a	18 (1.00) / 17 (1.00)	–	–	–	–	0.000 → 0.000
ghosting 20 / 40%	0.000 / n/a	36 (1.00) / 23 (1.00)	–	–	–	–	0.000 → 0.000
last rep. low-SNR 1× / 2×	1.000 / 0.017	– (0.00)	–	70	38	15	0.896 → 0.028
delivery: stage $k-1$ image	0.695 / 0.695	– (0.00)	1	141	86	50	0.613 → 0.045
delivery: no reconstructor	1.000 / 1.000	– (0.00)	1	70	38	15	0.896 → 0.028
population: 40 lowest-SNR	0.051 / 0.051	– (0.00)	–	–	–	–	0.044 → 0.044

G Additional experimental details

Domains and shared protocol. The ImageNet-pretrained cascade (ResNet-18, ResNet-50, ConvNeXt-Base) uses 800 calibration and 3125 deployment images from Imagenette (Howard, 2019). It exits at max-softmax ≥ 0.7, with ConvNeXt-Base as the full-budget reference. M4Raw (Lyu et al., 2023) provides repeated 0.3 T brain MRI acquisitions, with $65/59/59$ subjects assigned to train/calibration/test splits; disagreement occurs when the discrepancy in either contrast exceeds its tolerance relative to full acquisition (App. D). Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024) is evaluated on GSM8K (Cobbe et al., 2021) and MATH-500 (Lightman et al., 2024), with the plurality answer after 64 samples as the full-budget reference. A separate held-out LLM study is described in App. C.

Unless stated otherwise, $(\alpha, \delta) = (0.05, 0.05)$. The initial stopping rule is selected from fully completed calibration examples using a fixed-sequence Clopper–Pearson test. This is stopping-rule calibration; §5.2 separately calibrates audit-policy parameters to a target for additional normal-operation cost relative to the unaudited frozen rule.

The MRI and classifier shift studies use streams of length 3000, with a change at index 1500; the LLM shift studies use 6000 questions with the change also at index 1500. A label-free change monitor provides a comparison that requires no audits: it tests for changes in the means of two prefix statistics, the first stopping score and the proposed exit stage, using a 300-subject burn-in, a mean-change tolerance of 0.02 and total level δ across the two statistics. Audit alarms trigger latched full completion from the following subject; when both audit alarms are used, each runs at level $\delta/2$, so by a union bound their deployment-wide false-alarm probability under H_0 is at most δ.

Costs include natural deferrals, audits and fallback. We distinguish the proposed rules' disagreement risk from disagreement actually delivered after intervention, and identify delivery outcomes as whole-stream or post-change. The primary classifier study in §5.1 reports means over 30 paired streams; the budget-calibrated study in §5.2 uses 100 paired evaluation streams. Supporting studies retain the sample sizes and summaries specified in the appendix. Replay-based uncertainty is conditional on the fixed underlying subject or image pools and does not include uncertainty from drawing a new population sample. The budget and grid comparisons are follow-up studies whose designs were fixed before execution (App. F).

For the classifier, computation is measured by nominal multiply-adds accumulated over the stages actually executed and normalised by full completion on the same stream. When the early branch changes architecture or resolution, its corresponding nominal stage cost changes as well; later stages are charged whenever they are executed.

G.1 MRI acquisition anomalies

Table 23 and Figure 4 report seven acquisition anomalies, two delivery-path bugs and a population shift under the frozen offline-certified rule (App. D). The "clean ref." column uses the unchanged

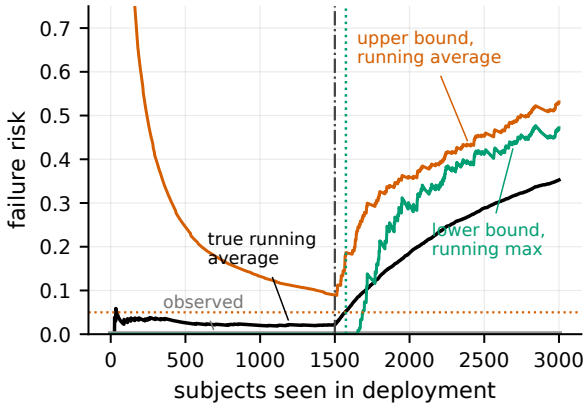

Figure 4: MRI delivery failure missed by score monitoring ($\bar{p} = 0.1$, one stream). Dash-dotted: failure begins; dotted: betting alarm. Black: running risk; grey: observed loss without audits; orange: upper bound on running risk; green: lower bound on maximum risk.

full-budget image only when the anomaly leaves the delivered prefix untouched; paired anomaly strengths with identical results share a row. *Disagreement within target:* permuting the repetition order or a 20% intensity drift leaves the risk below α $(0.000, 0.025)$; the alarms stay silent, the monitor fires on the drift after a median of 80 subjects despite the low disagreement. *Visible:* a 40% drift, a motion spike or a ghosting artefact make the rule defer every subject (risk zero, saving lost); the monitor fires within dozens of subjects, the alarms never. *Reference failure:* replacing the unobserved third repetition by a low-SNR draw leaves prefix, scores and defer rate unchanged, so the rule stops every subject and by (1) fails every one, yet the delivered image is unchanged and closer to the clean full-budget image than the completed one on 99–100% of units. The LR alarm fires after 70, 38 and 15 subjects at $\bar{p}{=}0.05, 0.1, 0.2$: the audits see that the prefix-to-full relation changed, not which image is right; the guardrail delivers the worse one. *Silent to the monitor, with excess disagreement:* two integration bugs in the delivery path leave prefix, score and defer rate untouched and the full-budget image intact: delivering the stage-$(k{-}1)$ reconstruction fails 0.70 of the stopped subjects, bypassing the reconstructor all of them. These bugs leave both monitored statistics unchanged; the LR alarm fires after 141, 86 and 50 subjects on the first and 70, 38, 15 on the second, and with the guardrail on the delivered failure rate after the change falls from 0.61 to 0.045 and from 0.90 to 0.03 at $\bar{p}{=}0.1$, at whole-stream net savings of 0.15 and 0.14 against 0.28. An integrity check recomputing $\mathcal{D}$ on the logged prefix catches both at once (Table 23), given a trusted copy of $\mathcal{D}$ and a delivery log; audits need only the recorded early output, and also see the reference side. *Population shift* (the 40 lowest-SNR subjects, risk 0.051): neither the monitor nor the audits fire within 1500 subjects, consistent with the near-boundary results in E8. Nine further splits repeat the picture (Table 38, App. G).

G.2 LLM SERVING CHANGES

On GSM8K the offline threshold selected on 400 fully sampled questions is $\lambda{=}0.2$: the rule stops after 3.6 samples on average with a population risk of 0.017 on the remaining 919 questions (Table 45, App. G). Four shifts are injected at question 1500 of a resampled stream (Table 44, App. G): a task shift to MATH-500 level ≥ 4 (risk 0.082); a serving bug that duplicates the first sample (0.100); a parser bug that blanks 30% of answers (0.161); and a batching bug that misroutes the previous request's answer (0.986). All four move the first-checkpoint score or the samples drawn; the monitor has lower-median delays of 192–557 questions. The misrouting bug is the LLM analogue of the MRI delivery bug: the LR alarm fires after 75, 38 and 15 questions (betting alarm 162, 86, 59), and with the guardrail on the delivered failure rate after the shift falls from 0.88 to 0.009 at $\bar{p}{=}0.1$ (Figure 14, App. G); the milder shifts take 528–2850 questions at $\bar{p}{=}0.2$. A pre-specified check on 1000 held-out GSM8K-train questions with fresh samples (App. C) reproduces the misrouting and parser delays $(38, 925)$; of ten freshly sampled serving configurations, including a regeneration control, nine meet the disagreement target (risks 0.001–0.043) and the audits stay silent on every one, while a 256-token

cap exceeds the target (risk 0.133) and the LR alarm catches it after 2120 questions at $\bar{p}$=0.1 (784 with escalation); the monitor fires on both the cap and model replacement; audit alarms trigger only on the cap, whose disagreement exceeds the target. In task terms (Table 47, App. G) the model replacement meets the disagreement target but raises the early output's error from 0.040 to 0.112 (full budget 0.083); delivered streams: Figure 11.

G.3 Cascade corruptions and integration bugs

The offline threshold selected on 800 images (App. G) exits when the current stage's max-softmax is at least 0.7: 63% of images exit after ResNet-18 and 35% run the whole cascade (FLOPs saving 59%, population risk 0.023; Table 40, App. G). Table 39 (App. G) injects nine input corruptions and two integration bugs. Every corruption is *visible*: the defer rate jumps from 0.35 to 0.43–0.86 and the monitor has lower-median delays of 55–425 images (all but a brightness offset); the rule risk stays at or below 0.0502 for all but pixelation, which raises it to 0.13 while the large model collapses: a reference failure, completing spoils 8% of the answers (Table 47); the alarms stay silent on every other corruption. The failures *missed by the tested monitors* are integration bugs in the label lookup table (an integrity check of the table against the logits would catch them, given a trusted copy): an off-by-one among the ten classes leaves stage 1's confidence, exit rate and label support untouched while 62% of images fail; swapping two classes of nearly equal frequency leaves even the label histogram nearly untouched (total variation 0.001), with 20% failing. The LR alarm fires after 128, 77 and 43 images for the off-by-one and after 1072, 576 and 284 for the swap (50%, 100%, 100% of streams), and with the guardrail on the delivered failure rate falls from 0.54 to 0.040 and from 0.18 to 0.076 at $\bar{p}$=0.1 (Figure 13, App. G).

Table 24: Component comparisons. Delays are in subjects at the stated audit rate; parentheses give detection fractions below one.

module	without	with	evidence
randomised audits (§4)	no audit: observed failure rate 0, no certificate	$\bar{p}$=0.1: bound 0.040 after 3000 subjects (313 audits)	E1
stationary betting CS vs drift-valid EB CS (Thm. 2)	EB (valid under drift): 0.068	betting (stationary): 0.040	E1, $\bar{p}$=0.1
prediction-powered increment (score as control variate, Eq. 3)	propensity weighting only: 113 (90%) audits to certify α (risk 0.017)	raw score: 122 (90%); isotonic: 118 (90%); alarms unchanged (LR ignores it; betting on the task shift 2387 (70%) without vs 2850 (70%) with)	E2, $\bar{p}$=0.1; E5, $\bar{p}$=0.2
LR alarm (Thm. 1) vs betting alarm (Cor. 1)	betting: 86 / 86 / 147 (catastrophic); 1050 (parser bug); 3547 (60%) (task shift)	LR: 38 / 38 / 77; 1520; 4500+ (40%)	E4, E5, E7, $\bar{p}$=0.1
anytime selector (Thm. 3) vs fixed-n re-checks	fixed-n on an iid audit subsample: 0.070; pooling every observed loss: 0.660	anytime selector: 0.005 (probability of ever deploying the unsafe candidate)	E3 synthetic, $\bar{p}$=0.2
audit escalation (Prop. 3) vs uniform / two-threshold at matched cost	uniform: 1178; two-threshold (1/3): 853	smooth escalation: 559 (restricted mean delay, risk 0.2)	E9, budget 0.05

Near-boundary study (E8). A synthetic cohort of two subject types with identical prefixes and acquisition costs (so that no label-free monitor can see the change), one failing and one not when stopped, with exact Bernoulli losses at a pre-change risk of 0.02 and post-change risks from 0.04 to 0.986; the change is placed at subject 1500 (a restart of the linear clock), 1501 and 1533, with 100 streams per cell, common random numbers across cells and a window of 6000 subjects; two no-change cells at risks 0.02 and $0.05 = \alpha$ over 7500 subjects measure false alarms (the subject-type identifier is bookkeeping, not part of the observable information: the monitors consume prefix scores and stages only). Observed false-alarm rates are at most 0.02 (LR) and 0.03 (betting) for every policy, 100-stream estimates rather than a precise check of the level. Under no change the paid rate stays at the floor at risk 0.02 but not at α: mean 0.062 for smooth escalation and 0.071 for the two-threshold

Table 25: Audit cost and detection on five changes exceeding the disagreement target (10 streams). Entries are medians except RMD (restricted mean delay). Post-change totals end at alarm or horizon H; parentheses give detection fractions below one. Extra cost is relative to no audits: MRI seconds, LLM samples, cascade GFLOPs.

shift	audit policy	before the change, per subject		LR alarm: through the alarm (capped at H)					betting alarm: through the alarm (capped at H)				
		overrides	extra cost	overrides	extra cost	failures	delay (frac.)	RMD	overrides	extra cost	failures	delay (frac.)	RMD
MRI delivery bug ($H = 1500$)	uniform 0.02	0.021	2.36	4	450	146	131	188	8	900	359	356	391
	uniform 0.05	0.050	5.72	4	480	69	70	80	8	930	160	162	170
	uniform 0.1	0.101	11.34	4	480	35	38	39	10	1140	80	86	96
	uniform 0.2	0.199	22.74	4	480	13	15	18	11	1290	55	66	65
	escalated, floor 0.02	0.021	2.38	4	480	91	92	108	13	1500	100	109	126
	escalated, floor 0.05	0.051	5.78	4	480	53	54	60	9	1050	57	63	70
	two-threshold, floor 0.02	0.021	2.36	4	450	144	131	168	13	1470	151	148	186
	two-threshold, floor 0.05	0.050	5.72	4	480	69	70	75	10	1110	73	76	85
LLM misrouting ($H = 4500$)	uniform 0.02	0.021	1.26	4	248	206	178	213	8	496	403	377	408
	uniform 0.05	0.050	3.04	4	248	75	75	83	8	496	152	162	172
	uniform 0.1	0.101	6.10	4	248	34	38	39	10	589	78	86	91
	uniform 0.2	0.199	12.04	4	248	12	15	18	11	682	49	59	61
	escalated, floor 0.02	0.021	1.26	4	248	90	92	95	13	798	95	109	111
	escalated, floor 0.05	0.052	3.16	4	248	49	49	58	9	558	52	57	66
	two-threshold, floor 0.02	0.021	1.26	4	248	172	164	175	13	806	178	180	190
	two-threshold, floor 0.05	0.050	3.04	4	248	75	75	78	9	558	81	88	88
cascade off-by-one ($H = 1500$)	uniform 0.02	0.014	0.27	5	98	204	299	408	8	156	323	520	639
	uniform 0.05	0.033	0.63	5	98	99	128	160	8	156	132	233	233
	uniform 0.1	0.064	1.24	7	127	51	77	88	10	195	85	147	156
	uniform 0.2	0.128	2.48	6	117	23	43	46	12	222	48	95	99
	escalated, floor 0.02	0.014	0.27	6	105	88	142	200	13	242	97	162	219
	escalated, floor 0.05	0.033	0.63	6	117	49	75	82	9	171	50	79	90
	two-threshold, floor 0.02	0.014	0.27	5	98	162	233	327	15	279	168	266	358
	two-threshold, floor 0.05	0.033	0.63	6	117	62	102	113	10	195	64	117	126
LLM parser bug ($H = 4500$)	uniform 0.02	0.021	1.26	87	5084	684	4336 (60%)	3795	89	5175	681	4298 (60%)	3987
	uniform 0.05	0.050	3.04	86	5045	242	1319	1650	75	4378	230	1282	1501
	uniform 0.1	0.101	6.10	172	10021	225	1520	1384	125	7286	188	1050	1188
	uniform 0.2	0.199	12.04	121	7060	71	545	683	118	6884	68	528	597
	escalated, floor 0.02	0.021	1.26	71	4164	123	661	1068	65	3862	123	666	1062
	escalated, floor 0.05	0.052	3.16	79	4576	80	518	863	76	4434	77	518	883
	two-threshold, floor 0.02	0.021	1.26	59	3444	249	1486	1732	113	6568	252	1514	1816
	two-threshold, floor 0.05	0.050	3.04	95	5567	145	1005	1112	74	4348	141	806	1094
cascade label swap ($H = 1500$)	uniform 0.02	0.014	0.27	18	341	280	1500+ (20%)	1399	19	359	280	1500+ (10%)	1464
	uniform 0.05	0.033	0.63	36	690	232	1072 (50%)	1094	34	651	185	976 (70%)	1063
	uniform 0.1	0.064	1.24	48	920	115	576	674	49	944	121	675	766
	uniform 0.2	0.128	2.48	50	961	45	284	447	64	1238	58	389	447
	escalated, floor 0.02	0.014	0.27	20	388	199	894 (60%)	1003	32	622	200	896 (60%)	1028
	escalated, floor 0.05	0.033	0.63	46	893	111	662	652	44	846	108	600	650
	two-threshold, floor 0.02	0.014	0.27	20	376	274	1500+ (40%)	1231	27	515	274	1500+ (40%)	1262
	two-threshold, floor 0.05	0.033	0.63	43	823	155	802 (90%)	870	48	918	154	795 (90%)	872

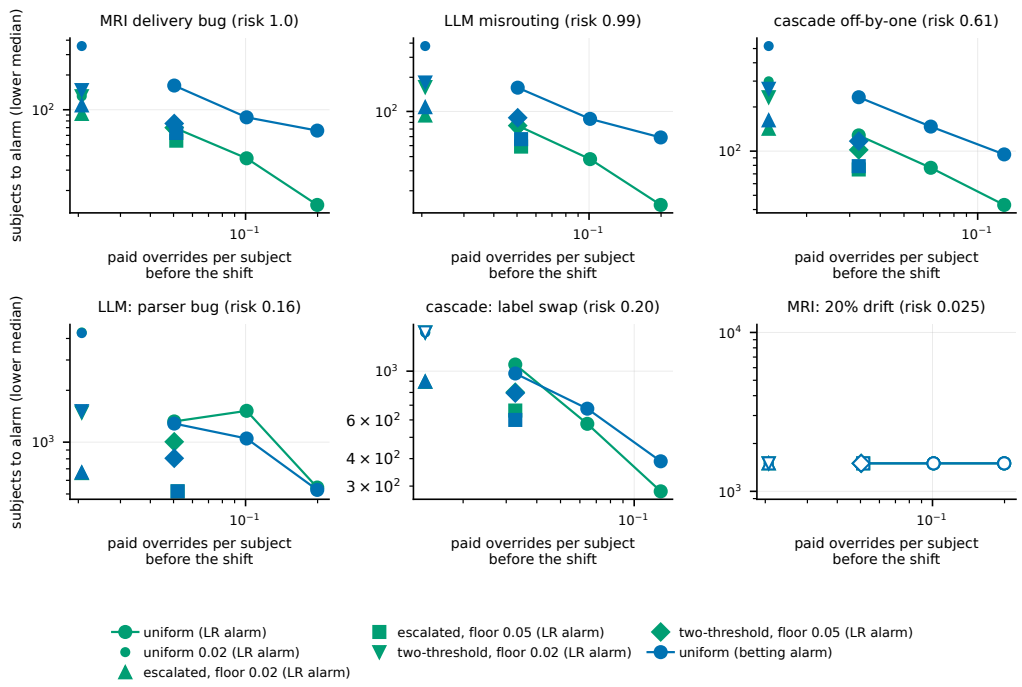

Figure 5: Delay versus paid audit rate before the change: lower medians over 10 streams. Hollow markers indicate censoring. Top: severe failures; bottom: milder changes, including MRI intensity drift below the target. Restricted means appear in Figure 6.

policy from a floor of 0.05 (medians 0.053 and 0.051), reflecting evidence fluctuations below the alarm threshold. Below $r = 0.0724$ no alternative has positive drift and Theorem 1 gives no bound; every lower median is censored at 6000 below $r = 0.08$, with detection fractions rising from 0.01 at $r = 0.055$ to 0.21 at 0.072 (a restarted mixture can cross in a finite window without positive drift); at $r = 0.08$ the bound (22577) exceeds the window and uniform 0.2 detects on 56% of streams. From $r = 0.1$ on, the lower medians at $\bar{p}$=0.1 (3531, 578, 141, 35 subjects at $r = 0.1, 0.2, 0.5, 0.986$; restricted means 3688, 657, 148, 40) are below the change-relative bound on the expected delay (6242, 972, 177, 53) at all three change times, a comparison of magnitude only, since censored means do not verify the full expectation. The bound for a change at 1533 adds the wait to the next clean linear restart (17 subjects) and a smaller restart weight (71 against 53 at $r = 0.986$); the observed medians move from 35 to 54 at $r = 0.986$ and from 3531 to 3689 at $r = 0.1$, differences the theorem does not bound. At approximately matched pre-shift cost (0.051 paid overrides per subject), escalation from floor 0.05 has a lower median below the default two-threshold policy and below uniform 0.05 at every risk from 0.1 up, and below uniform 0.1 (paid 0.100) up to $r = 0.5$ (restricted means 149 against 148 at $r = 0.5$). The switch-on level matters: at $1/6$ the two-threshold policy is faster than the smooth schedule at $r = 0.1$ (2112 against 2269; restricted means 2771 against 2830) and slower at 0.5 (151 against 134), while $1/2$ gives 2972 and 203 (Table 27); the smooth schedule has its own choices (exponent, evidence source, ceiling), so E8 alone establishes an advantage over uniform audits and over the default on–off controller; E9 below calibrates a common cost target and varies both families. Table 25 prices every policy through the alarm on the five changes exceeding the disagreement target in the three domains, and Figure 7 shows why the default two-threshold controller rarely helps on a near-certain shift. The switch-on and alarm levels are crossed on the same audited failure, leaving no time for the two-threshold controller to increase observation before detection.

Budget-calibrated policy comparison (E9). Specified before the run (protocol in the repository): on the E8 cohort, seven policies (uniform; smooth escalation with exponent $\gamma \in \{0.25, 0.5, 1\}$; two-threshold control with switch-on level $\{1/6, 1/3, 1/2\}$, off at zero; ceiling 0.5) have their audit-rate or floor parameter calibrated on separate seeds so that the paid override rate until the first alarm under the null at risk α targets a budget of 0.05 or 0.1, and are then evaluated on 100 fresh streams per cell

Table 26: E8: LR lower-median delay (detection fraction below one; +: censored at 6000). Headers give paid rates before the change. Bounds use Theorem 1 at $\bar{p} = 0.1$; ∞ means no positive drift. Null rows report false-alarm rates over 7500 subjects.

post-change risk	change time	LR bound ($\bar{p}$=0.1)	LR alarm: lower-median delay (fraction detected when below one; +: censored at $H = 6000$); null rows: false-alarm rate LR / betting over the whole horizon									
			uniform 0.02 @ 0.019	uniform 0.05 @ 0.050	uniform 0.1 @ 0.100	uniform 0.2 @ 0.199	escalated 0.02 @ 0.020	escalated 0.05 @ 0.051	two-threshold 0.02 @ 0.020	two-threshold 0.05 @ 0.051	two-thr. 0.05, on 1/6 @ 0.051	two-thr. 0.05, on 1/2 @ 0.050
null $r = 0.02$	–	–	0.00 / 0.00 @ 0.020	0.00 / 0.00 @ 0.050	0.00 / 0.00 @ 0.100	0.00 / 0.00 @ 0.200	0.00 / 0.00 @ 0.020	0.00 / 0.00 @ 0.050	0.00 / 0.00 @ 0.020	0.00 / 0.00 @ 0.050	–	–
null $r = 0.05$	–	–	0.00 / 0.00 @ 0.020	0.02 / 0.01 @ 0.050	0.01 / 0.03 @ 0.100	0.01 / 0.00 @ 0.200	0.02 / 0.02 @ 0.022	0.02 / 0.02 @ 0.053	0.01 / 0.00 @ 0.021	0.02 / 0.02 @ 0.051	–	–
0.04	1500	∞	–	–	–	–	–	–	–	–	–	–
0.04	1501	∞	–	–	–	–	–	–	–	–	–	–
0.04	1533	∞	–	–	–	–	–	–	–	–	–	–
0.055	1500	∞	–	–	–	–	6000+ (1%)	6000+ (1%)	–	6000+ (1%)	–	–
0.055	1501	∞	–	–	–	–	6000+ (1%)	6000+ (1%)	–	6000+ (1%)	–	–
0.055	1533	∞	–	–	–	–	–	6000+ (1%)	–	6000+ (1%)	–	–
0.06	1500	∞	6000+ (1%)	–	–	6000+ (1%)	6000+ (4%)	6000+ (2%)	6000+ (1%)	6000+ (4%)	–	–
0.06	1501	∞	6000+ (1%)	–	–	6000+ (1%)	6000+ (4%)	6000+ (2%)	6000+ (1%)	6000+ (4%)	–	–
0.06	1533	∞	6000+ (1%)	–	–	6000+ (1%)	6000+ (3%)	6000+ (2%)	6000+ (1%)	6000+ (4%)	–	–
0.07	1500	∞	6000+ (3%)	6000+ (3%)	6000+ (4%)	6000+ (13%)	6000+ (10%)	6000+ (14%)	6000+ (8%)	6000+ (13%)	–	–
0.07	1501	∞	6000+ (3%)	6000+ (3%)	6000+ (4%)	6000+ (13%)	6000+ (10%)	6000+ (14%)	6000+ (8%)	6000+ (13%)	–	–
0.07	1533	∞	6000+ (3%)	6000+ (3%)	6000+ (4%)	6000+ (13%)	6000+ (9%)	6000+ (14%)	6000+ (8%)	6000+ (12%)	–	–
0.072	1500	∞	6000+ (3%)	6000+ (3%)	6000+ (5%)	6000+ (21%)	6000+ (12%)	6000+ (18%)	6000+ (11%)	6000+ (18%)	–	–
0.072	1501	∞	6000+ (3%)	6000+ (3%)	6000+ (5%)	6000+ (21%)	6000+ (12%)	6000+ (18%)	6000+ (11%)	6000+ (18%)	–	–
0.072	1533	∞	6000+ (3%)	6000+ (3%)	6000+ (5%)	6000+ (21%)	6000+ (11%)	6000+ (18%)	6000+ (11%)	6000+ (17%)	–	–
0.073	1500	268886	6000+ (4%)	6000+ (4%)	6000+ (8%)	6000+ (24%)	6000+ (13%)	6000+ (20%)	6000+ (14%)	6000+ (21%)	–	–
0.073	1501	270239	6000+ (4%)	6000+ (4%)	6000+ (8%)	6000+ (24%)	6000+ (13%)	6000+ (20%)	6000+ (14%)	6000+ (21%)	–	–
0.073	1533	270207	6000+ (4%)	6000+ (4%)	6000+ (8%)	6000+ (24%)	6000+ (12%)	6000+ (20%)	6000+ (14%)	6000+ (20%)	–	–
0.08	1500	22577	6000+ (5%)	6000+ (9%)	6000+ (26%)	5294 (56%)	6000+ (28%)	6000+ (42%)	6000+ (24%)	5996 (50%)	–	–
0.08	1501	22736	6000+ (5%)	6000+ (9%)	6000+ (26%)	5293 (56%)	6000+ (28%)	6000+ (42%)	6000+ (24%)	5995 (50%)	–	–
0.08	1533	22704	6000+ (4%)	6000+ (9%)	6000+ (25%)	5261 (56%)	6000+ (27%)	6000+ (40%)	6000+ (25%)	6000+ (49%)	–	–
0.1	1500	6242	6000+ (16%)	5452 (56%)	3531 (83%)	2019 (99%)	3755 (76%)	2269 (84%)	4871 (58%)	2446 (80%)	2112 (83%)	2972 (75%)
0.1	1501	6321	6000+ (16%)	5451 (56%)	3530 (83%)	2018 (99%)	3754 (76%)	2268 (84%)	4870 (58%)	2445 (80%)	–	–
0.1	1533	6289	6000+ (15%)	5419 (58%)	3689 (83%)	1986 (99%)	3742 (76%)	2236 (84%)	4838 (58%)	2413 (79%)	–	–
0.2	1500	972	2550 (98%)	1063	578	305	835	441	1405 (99%)	519	–	–
0.2	1501	1026	2549 (98%)	1062	577	304	834	440	1404 (99%)	518	–	–
0.2	1533	994	2517 (99%)	1046	572	292	824	429	1372 (99%)	500	–	–
0.5	1500	177	530	255	141	75	217	134	329	189	151	203
0.5	1501	227	529	255	142	75	219	136	332	188	–	–
0.5	1533	195	500	239	143	80	206	140	320	174	–	–
0.986	1500	53	201	75	35	18	95	51	156	73	–	–
0.986	1501	103	200	75	36	20	94	53	155	72	–	–
0.986	1533	71	209	99	54	36	94	59	158	79	–	–

Table 27: E8 at change time 1500: detection fraction, restricted mean delay and failures before alarm, all within 6000 subjects. Two-threshold variants use switch levels 1/6 and 1/2.

alarm	risk	fraction detected / restricted mean delay at $H = 6000$ / failures delivered before the alarm (change at 1500)									
		uniform 0.02	uniform 0.05	uniform 0.1	uniform 0.2	escalated 0.02	escalated 0.05	two-threshold 0.02	two-threshold 0.05	two-thr. 0.05, on 1/6	two-thr. 0.05, on 1/2
LR	0.08	5% / 5908 / 470	9% / 5835 / 452	26% / 5394 / 421	56% / 4456 / 340	28% / 5203 / 448	42% / 4806 / 417	24% / 5433 / 456	50% / 4748 / 408	–	–
LR	0.1	16% / 5666 / 585	56% / 4789 / 518	83% / 3688 / 312	99% / 2252 / 156	76% / 3710 / 357	84% / 2830 / 194	58% / 4375 / 476	80% / 3047 / 220	83% / 2771 / 182	75% / 3454 / 271
LR	0.2	98% / 2713 / 506	100% / 1178 / 204	100% / 657 / 107	100% / 345 / 49	100% / 1072 / 156	100% / 547 / 78	99% / 1598 / 279	100% / 666 / 100	–	–
LR	0.5	100% / 620 / 258	100% / 270 / 121	100% / 148 / 66	100% / 79 / 31	100% / 266 / 100	100% / 149 / 61	100% / 431 / 156	100% / 193 / 85	100% / 163 / 68	100% / 215 / 92
LR	0.986	100% / 204 / 194	100% / 79 / 70	100% / 40 / 31	100% / 20 / 14	100% / 105 / 90	100% / 55 / 46	100% / 163 / 151	100% / 74 / 68	–	–
betting	0.08	1% / 5975 / 472	14% / 5747 / 452	36% / 5012 / 418	59% / 4126 / 278	30% / 5127 / 447	48% / 4498 / 406	24% / 5428 / 456	53% / 4520 / 407	–	–
betting	0.1	7% / 5839 / 588	56% / 4645 / 504	88% / 3161 / 240	99% / 1890 / 125	76% / 3663 / 359	85% / 2693 / 178	56% / 4447 / 484	80% / 2989 / 216	84% / 2681 / 178	74% / 3427 / 268
betting	0.2	84% / 3647 / 657	100% / 1232 / 202	100% / 642 / 108	100% / 364 / 56	100% / 1116 / 162	100% / 546 / 79	99% / 1689 / 288	100% / 685 / 100	–	–
betting	0.5	100% / 906 / 398	100% / 357 / 164	100% / 206 / 86	100% / 123 / 46	100% / 296 / 108	100% / 162 / 64	100% / 468 / 162	100% / 208 / 88	100% / 179 / 72	100% / 229 / 93
betting	0.986	100% / 397 / 374	100% / 167 / 158	100% / 97 / 79	100% / 60 / 48	100% / 125 / 97	100% / 67 / 52	100% / 183 / 159	100% / 86 / 72	–	–

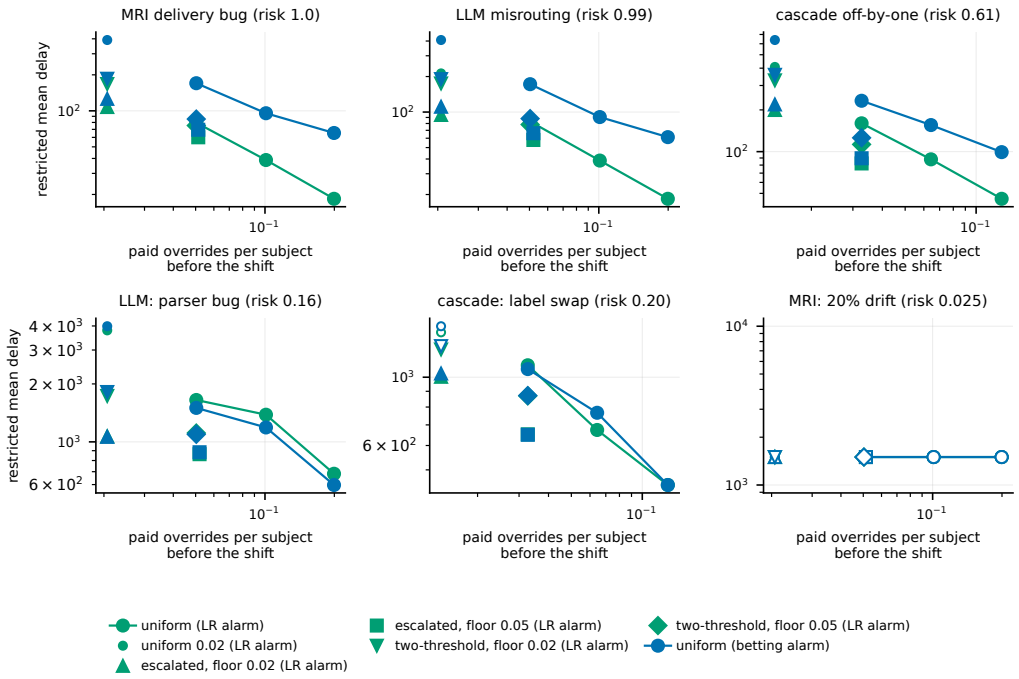

Figure 6: Restricted mean delays for the policies and streams in Figure 5, truncated at each domain's observation horizon.

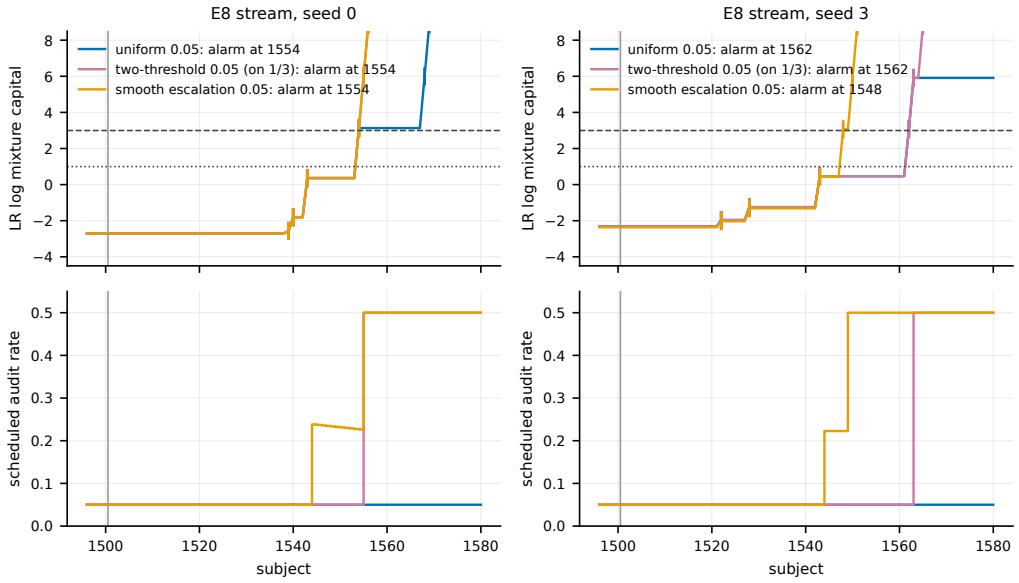

Figure 7: Two E8 streams with risk $0.02 \to 0.986$ at subject 1500. Top: LR log capital; bottom: audit rate. Ticks: audited failures; dotted/dashed lines: switch-on/alarm thresholds.

at risks 0.1, 0.2, 0.5, 0.986 and change times 1500, 1501, 1533, with the restricted mean delay of the LR alarm at $H = 6000$ as the primary endpoint at budget 0.05, change 1500 and the three moderate risks, compared pairwise over common random numbers with 95% bootstrap intervals. Calibrated rates or floors at budget 0.05: uniform 0.050, smooth 0.043–0.047, two-threshold 0.026–0.040; realised false-alarm rates in the null cells stay at or below 0.05. Result (Table 28, Figure 9): the

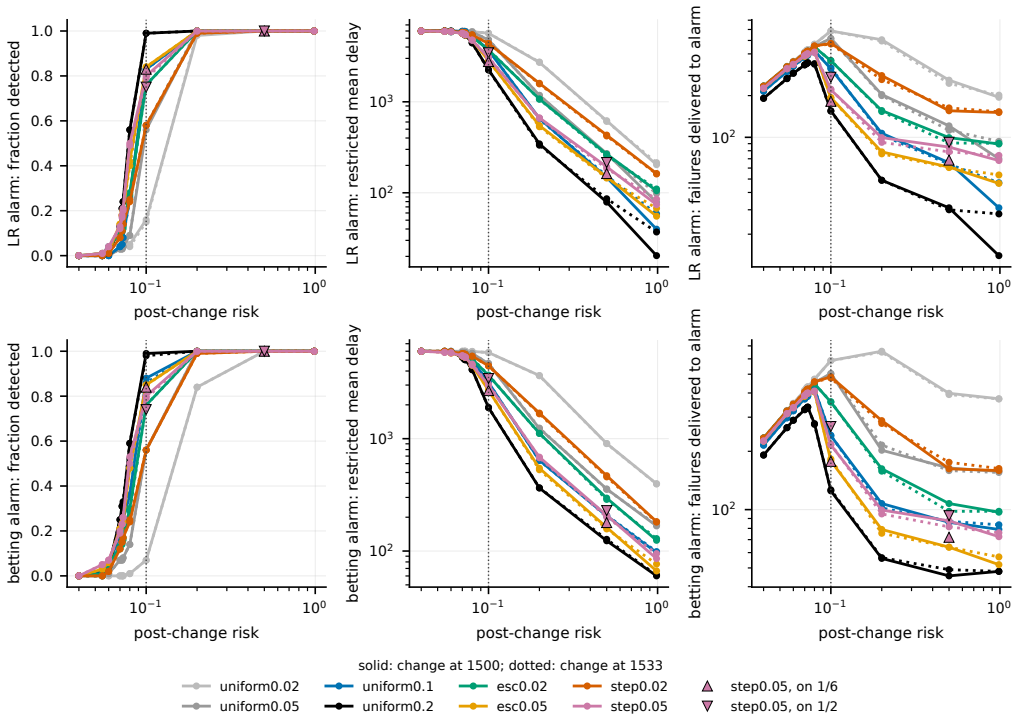

Figure 8: Near-boundary detection (E8): 6000-subject window, LR alarm top and betting alarm bottom. Solid/dotted curves: changes at $1500/1533$. Triangles: two-threshold switch levels $1/6$ and $1/2$. Vertical line: 2α.

paper's schedule ($\gamma = 0.5$) is faster than the best two-threshold level ($1/3$ at every risk) by 573, 294 and 83 subjects at $r = 0.1, 0.2, 0.5$ (intervals $[-860, -273]$, $[-378, -218]$, $[-100, -66]$; Table 29) and than uniform audits by 2062, 619 and 111, with the same ordering at the other two change times and for the betting alarm, and fewer failures delivered before the alarm (193, 76, 66 against 246, 122, 96 and 518, 204, 121). At budget 0.1 the $1/6$ switch is faster at $r = 0.1$ ($+174$, $[66, 288]$) and slower at 0.2 and 0.5; $\gamma = 0.25$ is slightly faster than 0.5 everywhere ($+94$ $[-39, 241]$, $+25$ $[0, 52]$, $+10$ $[6, 15]$). Because calibration targets the null at α, the on–off policies pay less than uniform at the safe null (0.033–0.041 against 0.050); the equal-floor comparison of E8 is the safe-null matching, where the $1/6$ switch wins at $r = 0.1$ only.

Joint policy (E12). The monitor of §5 and the audits can be combined without touching any guarantee: the monitor's two betting tests have a log capital, which over their alarm level and clipped to $[0, 1]$ is an evidence statistic like the alarms' own; read one subject behind, it is $\mathcal{F}_{i-1}$-measurable, so Proposition 3 applies with it (or with the maximum of it and the alarms' evidence) as e_{i-1}. Four policies with floor 0.05 and ceiling 0.5 run on the shift conditions of the three domains (10 streams each, the MRI streams those of Table 23): uniform audits, alarm-driven escalation (the paper's), monitor-driven escalation and their maximum. We evaluate two deployments: Table 30 the frozen rule with the LR alarm at level δ (detection: alarm delay, mean assigned audit probability among the subjects the rule stops after the change, failures delivered before the alarm), Table 31 the guardrail on with both alarms at $\delta/2$ (intervention: the guardrail's own alarm delay, the same assigned probability among the subjects it still stops, delivered disagreement after the change and whole-stream net saving with every audit and the fallback charged). The audit-probability denominator excludes subjects completed by fallback. E12 cascade streams use different seeds from Table 39. Assigned audit probability determines sampling; net saving measures its realised cost. On a change the monitor sees and that exceeds the disagreement target, the monitor's early evidence speeds auditing: the token cap's LR alarm moves from a lower median of 2197 questions (uniform, 90% of streams) and 641 (alarm-driven) to 333 (every stream); paired over the ten streams, undetected streams censored at the 4500-question window, the joint policy's restricted mean delay is 1062 questions

Table 28: E9: calibrated rates or floors, realised null costs and false-alarm rates (7500 subjects), followed by LR detection results (change at 1500, horizon 6000). RMD is restricted mean delay; failures are medians.

policy	calibrated floor	realised cost: null at α / at 0.02	false alarms LR / betting	LR alarm, change at 1500: fraction detected / restricted mean delay at $H = 6000$ / failures before the alarm			
				$r = 0.1$	$r = 0.2$	$r = 0.5$	$r = 0.986$
budget $B = 0.05$ (paid overrides per subject under the null at α)							
uniform	0.0500	0.0506 / 0.0503	0.02 / 0.01	56% / 4789 / 518	100% / 1178 / 204	100% / 270 / 121	100% / 79 / 70
smooth γ=0.25	0.0427	0.0579 / 0.0435	0.02 / 0.02	90% / 2634 / 169	100% / 534 / 70	100% / 149 / 58	100% / 62 / 52
smooth γ=0.5 (paper)	0.0425	0.0549 / 0.0432	0.03 / 0.02	90% / 2728 / 193	100% / 559 / 76	100% / 159 / 66	100% / 65 / 52
smooth γ=1	0.0470	0.0581 / 0.0475	0.03 / 0.02	78% / 3315 / 256	100% / 606 / 91	100% / 180 / 78	100% / 65 / 55
two-threshold, on 1/6	0.0260	0.0535 / 0.0328	0.01 / 0.01	79% / 3358 / 277	100% / 1021 / 159	100% / 260 / 98	100% / 109 / 93
two-threshold, on 1/3 (paper)	0.0380	0.0509 / 0.0384	0.04 / 0.01	76% / 3301 / 246	100% / 853 / 122	100% / 242 / 96	100% / 97 / 83
two-threshold, on 1/2	0.0403	0.0479 / 0.0407	0.05 / 0.01	74% / 3659 / 344	100% / 934 / 148	100% / 255 / 105	100% / 100 / 86
budget $B = 0.1$ (paid overrides per subject under the null at α)							
uniform	0.1000	0.1004 / 0.1003	0.01 / 0.03	83% / 3688 / 312	100% / 657 / 107	100% / 148 / 66	100% / 40 / 31
smooth γ=0.25	0.0923	0.1042 / 0.0931	0.01 / 0.01	94% / 2209 / 141	100% / 395 / 57	100% / 105 / 42	100% / 32 / 23
smooth γ=0.5 (paper)	0.0946	0.1019 / 0.0952	0.01 / 0.03	94% / 2369 / 154	100% / 399 / 60	100% / 111 / 47	100% / 34 / 24
smooth γ=1	0.0969	0.1029 / 0.0974	0.01 / 0.02	93% / 2568 / 173	100% / 442 / 67	100% / 116 / 49	100% / 37 / 29
two-threshold, on 1/6	0.0878	0.1075 / 0.0881	0.02 / 0.02	92% / 2195 / 140	100% / 431 / 56	100% / 121 / 46	100% / 42 / 33
two-threshold, on 1/3 (paper)	0.0816	0.0963 / 0.0820	0.05 / 0.03	93% / 2439 / 155	100% / 500 / 72	100% / 141 / 58	100% / 49 / 41
two-threshold, on 1/2	0.0908	0.1002 / 0.0912	0.03 / 0.03	90% / 2682 / 196	100% / 507 / 74	100% / 138 / 58	100% / 43 / 34

Table 29: E9: paired LR restricted-mean-delay differences, smooth escalation ($\gamma = 0.5$) minus each comparator, with 95% bootstrap intervals. Change at 1500; 100 shared seeds (99 for switch level 1/6). Negative values favour smooth escalation.

budget	policy	restricted mean delay of smooth γ=0.5 minus the policy (LR alarm, change at 1500; paired over seeds, 95% bootstrap interval)			
		$r = 0.1$	$r = 0.2$	$r = 0.5$	$r = 0.986$
$B = 0.05$	uniform	-2062 [-2407, -1718]	-619 [-711, -528]	-111 [-131, -92]	-14 [-20, -8]
$B = 0.05$	smooth γ=0.25	+94 [-39, 241]	+25 [0, 52]	+10 [6, 15]	+2 [1, 4]
$B = 0.05$	smooth γ=1	-588 [-851, -376]	-48 [-90, -6]	-21 [-30, -14]	-0 [-4, 4]
$B = 0.05$	two-threshold, on 1/6	-612 [-962, -291]	-463 [-621, -320]	-100 [-125, -77]	-44 [-55, -34]
$B = 0.05$	two-threshold, on 1/3 (paper)	-573 [-860, -273]	-294 [-378, -218]	-83 [-100, -66]	-33 [-40, -26]
$B = 0.05$	two-threshold, on 1/2	-932 [-1239, -644]	-376 [-464, -299]	-96 [-112, -79]	-35 [-42, -29]
$B = 0.1$	uniform	-1319 [-1589, -1071]	-257 [-319, -203]	-37 [-45, -29]	-6 [-9, -4]
$B = 0.1$	smooth γ=0.25	+160 [35, 283]	+4 [-13, 21]	+6 [2, 10]	+2 [1, 3]
$B = 0.1$	smooth γ=1	-200 [-321, -87]	-42 [-63, -24]	-5 [-8, -2]	-4 [-5, -2]
$B = 0.1$	two-threshold, on 1/6	+174 [66, 288]	-32 [-66, -1]	-10 [-16, -4]	-9 [-11, -6]
$B = 0.1$	two-threshold, on 1/3 (paper)	-70 [-278, 136]	-100 [-142, -63]	-30 [-38, -23]	-16 [-19, -12]
$B = 0.1$	two-threshold, on 1/2	-313 [-535, -122]	-107 [-142, -73]	-27 [-34, -20]	-10 [-12, -8]

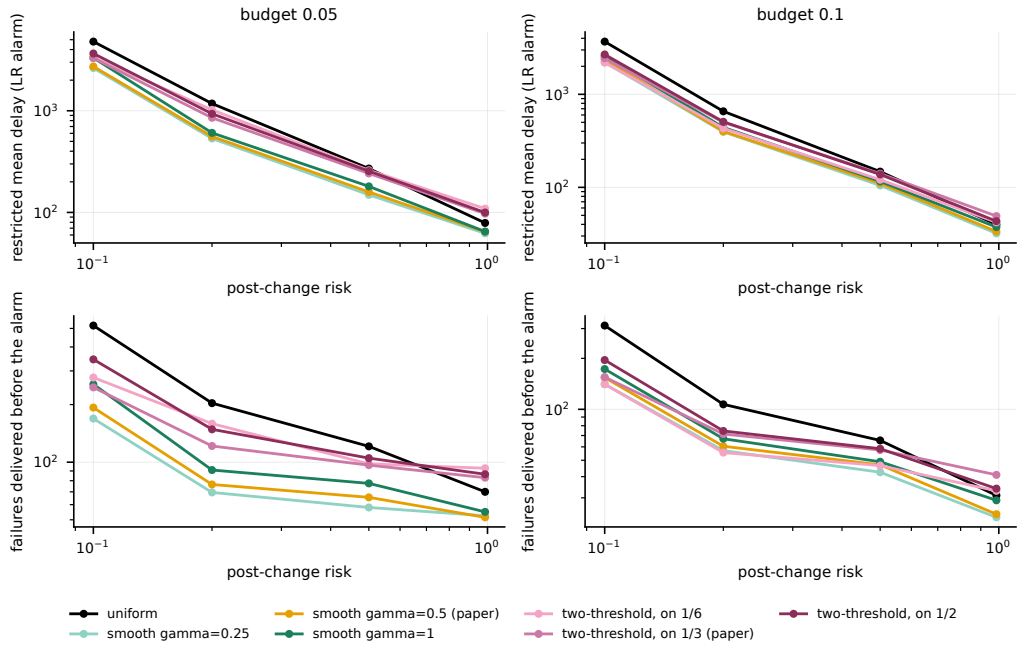

Figure 9: E9: LR restricted mean delay (top) and failures before alarm (bottom), change at 1500. Seven schedules share a calibrated null-cost target of 0.05 (left) or 0.1 (right).

below alarm-driven escalation's (95% bootstrap interval $[-1802, -380]$) and 2129 below uniform audits' ($[-2829, -1440]$), with 28 failures delivered before the alarm instead of 94 and 307; the price is the audit and fallback cost, a whole-stream net saving with the guardrail of 0.256 against 0.319 and 0.601, and the delivered disagreement after the change falls to 0.006 against 0.020 and 0.077. On the delivery bugs, which the monitor cannot see, monitor-driven escalation is uniform auditing and the joint policy is alarm-driven escalation. On visible changes within the disagreement target (the 20% drift, the replacement with the general model, most cascade corruptions) the rate rises to the ceiling for the rest of the stream and no fallback follows: the price of running both is the audit cost on visible changes within the target (the model replacement's net saving falls from 0.873 to 0.572), that alarm-driven escalation avoids. The cascade's pixelation is the reference-failure stress test of this study: the joint policy finds the disagreement after 290 images against 788 and undetected on 70% of streams, but the full budget's task error under pixelation is 0.916 against 0.838 for the early output (Table 47), so the faster fallback delivers the worse output sooner; faster detection of disagreement is a better intervention only where the reference is worth reaching.

End-to-end trade-off (E13). The question behind Tables 30 and 31 is whether the joint policy improves the trade-off between delivered disagreement and acquisition cost. E13 measures both over whole streams. For every policy of a grid fixed before the run (do nothing; never stop; a change monitor that falls back to full acquisition as soon as it fires and never audits; uniform audits at $p \in \{0.02, 0.05, 0.1, 0.2\}$; alarm-driven and joint escalation with floors 0.02, 0.05, 0.1 and ceiling 0.5; every audited policy with the guardrail on and both alarms at $\delta/2$), the streams of E12 give the cost (1− net saving with every audit, deferral and fallback charged), the delivered disagreement (fraction of the T subjects whose delivered output differs from the full-budget output; zero for audits, natural deferrals and completion after fallback) and, where labels exist, the delivered task error. Every grid point is reported (Figures 10 and 11; Tables 32, 33 and 34); the floor-0.05 policies replay the deployments of Table 31 (identical whole-stream cost; disagreement counted over the whole stream here, after the change there). The integrity check of App. G.1, which needs a trusted copy of $\mathcal{D}$, is not among the compared policies. The monitor fallback responds to statistical change. We report outcomes by the pre-specified condition classes.

Table 30: E12 detection with a frozen rule: 10 streams per condition, change at 1500, LR alarm at level δ. Audit floor 0.05, ceiling 0.5. Entries show lower-median delay (detection fraction), mean assigned audit probability and median failures before alarm.

condition	rule risk after	monitor delay	uniform 0.05			alarm-driven escalation			monitor-driven escalation			joint (max of both)		
			LR delay	assigned p	failures	LR delay	assigned p	failures	LR delay	assigned p	failures	LR delay	assigned p	failures
MRI clean control	0.017	–	–	0.05	26	–	0.05	26	–	0.05	26	–	0.05	26
MRI order permuted	0.000	–	–	0.05	0	–	0.05	0	–	0.05	0	–	0.05	0
MRI drift $+20\%$	0.025	80	–	0.05	34	–	0.05	34	–	0.48	19	–	0.48	19
MRI drift $+40\%$	0.000	36	–	–	0	–	–	0	–	–	0	–	–	0
MRI motion 8 px	0.000	17	–	–	0	–	–	0	–	–	0	–	–	0
MRI ghosting 40%	0.000	23	–	–	0	–	–	0	–	–	0	–	–	0
MRI low-SNR $2\times$ (reference failure)	1.000	–	70	0.05	69	54	0.48	53	70	0.05	69	54	0.48	53
MRI delivery: stage $k-1$ image	0.695	–	141	0.05	89	88	0.47	60	141	0.05	89	88	0.47	60
MRI delivery: no reconstructor	1.000	–	70	0.05	69	54	0.48	53	70	0.05	69	54	0.48	53
MRI population: lowest-SNR	0.051	–	–	0.05	73	–	0.05	73	–	0.05	73	–	0.05	73
LLM regeneration (no change)	0.005	–	–	0.05	21	–	0.05	21	–	0.05	21	–	0.05	21
LLM misrouted to the general model	0.043	82	–	0.05	183	–	0.05	183	–	0.49	95	–	0.49	95
LLM token cap $1024 \rightarrow 256$	0.133	67	2197 (90%)	0.05	308	641	0.45	94	333	0.49	29	333	0.49	29
cascade gauss noise 0.08	0.028	338	–	0.05	37	–	0.05	37	–	0.41	22	–	0.41	22
cascade gauss noise 0.16	0.050	69	–	0.05	72	–	0.05	72	–	0.48	39	–	0.48	39
cascade blur 2.0	0.029	181	–	0.05	41	–	0.05	41	–	0.45	26	–	0.45	26
cascade blur 4.0	0.023	58	–	0.05	33	–	0.05	33	–	0.49	18	–	0.49	18
cascade jpeg 10	0.027	304	–	0.05	40	–	0.05	40	–	0.43	26	–	0.43	26
cascade jpeg 5	0.033	63	–	0.05	48	–	0.05	48	–	0.48	27	–	0.48	27
cascade contrast 0.3	0.032	390	–	0.05	49	–	0.05	49	–	0.39	32	–	0.39	32
cascade brightness 0.3	0.027	–	–	0.05	40	–	0.05	40	–	0.05	40	–	0.05	40
cascade pixelate 0.25 (reference failure)	0.134	71	1500+ (30%)	0.05	188	788 (80%)	0.29	108	290	0.48	25	290	0.48	25
cascade lookup off-by-one	0.615	–	167	0.05	115	84	0.48	46	167	0.05	115	84	0.48	46
cascade lookup swap	0.196	–	796 (90%)	0.05	159	446	0.38	86	796 (90%)	0.05	159	446	0.38	86

Table 31: E12 intervention on the streams of Table 30, with both alarms at $\delta/2$. Entries show lower-median delay (detection fraction), mean audit probability among proposed early stops, post-change delivered disagreement and whole-stream net saving. All acquisition is charged.

condition	rule risk	uniform 0.05				alarm-driven escalation				monitor-driven escalation				joint (max of both)			
	after	alarm	assigned p	delivered	saving	alarm	assigned p	delivered	saving	alarm	assigned p	delivered	saving	alarm	assigned p	delivered	saving
MRI clean control	0.017	–	0.05	0.017	0.295	–	0.05	0.017	0.295	–	0.05	0.017	0.295	–	0.05	0.017	0.295
MRI order permuted	0.000	–	0.05	0.000	0.279	–	0.05	0.000	0.279	–	0.05	0.000	0.278	–	0.05	0.000	0.278
MRI drift $+20\%$	0.025	–	0.05	0.023	0.243	–	0.05	0.023	0.243	–	0.48	0.013	0.199	–	0.48	0.013	0.199
MRI drift $+40\%$	0.000	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145
MRI motion 8 px	0.000	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145
MRI ghosting 40%	0.000	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145	–	–	0.000	0.145
MRI low-SNR $2\times$ (reference failure)	1.000	89	0.05	0.058	0.154	55	0.09	0.036	0.150	89	0.05	0.058	0.154	55	0.09	0.036	0.150
MRI delivery: stage $k-1$ image	0.695	141	0.05	0.062	0.160	88	0.09	0.040	0.153	141	0.05	0.062	0.160	88	0.09	0.040	0.153
MRI delivery: no reconstructor	1.000	89	0.05	0.058	0.154	55	0.09	0.036	0.150	89	0.05	0.058	0.154	55	0.09	0.036	0.150
MRI population: lowest-SNR	0.051	–	0.05	0.049	0.285	–	0.05	0.049	0.285	–	0.05	0.049	0.285	–	0.05	0.049	0.285
LLM regeneration (no change)	0.005	–	0.05	0.005	0.916	–	0.05	0.005	0.916	–	0.05	0.005	0.916	–	0.05	0.005	0.916
LLM misrouted to the general model	0.043	–	0.05	0.041	0.873	–	0.05	0.041	0.873	–	0.49	0.021	0.572	–	0.49	0.021	0.572
LLM token cap $1024 \to 256$	0.133	2443 (90%)	0.05	0.077	0.601	639	0.13	0.020	0.319	322	0.42	0.006	0.256	324	0.42	0.006	0.256
cascade gauss noise 0.08	0.028	–	0.05	0.025	0.457	–	0.05	0.025	0.457	–	0.41	0.014	0.385	–	0.41	0.014	0.384
cascade gauss noise 0.16	0.050	–	0.05	0.048	0.360	–	0.05	0.048	0.360	–	0.48	0.026	0.324	–	0.48	0.026	0.324
cascade blur 2.0	0.029	–	0.05	0.027	0.430	–	0.05	0.027	0.430	–	0.45	0.017	0.366	–	0.45	0.017	0.366
cascade blur 4.0	0.023	–	0.05	0.022	0.340	–	0.05	0.022	0.340	–	0.49	0.012	0.311	–	0.49	0.012	0.311
cascade jpeg 10	0.027	–	0.05	0.027	0.444	–	0.05	0.027	0.444	–	0.43	0.017	0.377	–	0.43	0.017	0.377
cascade jpeg 5	0.033	–	0.05	0.032	0.341	–	0.05	0.032	0.341	–	0.48	0.018	0.313	–	0.48	0.018	0.313
cascade contrast 0.3	0.032	–	0.05	0.032	0.463	–	0.05	0.032	0.463	–	0.39	0.021	0.395	–	0.39	0.021	0.394
cascade brightness 0.3	0.027	–	0.05	0.026	0.520	–	0.05	0.026	0.519	–	0.05	0.026	0.520	–	0.05	0.026	0.519
cascade pixelate 0.25 (reference failure)	0.134	1500+ (40%)	0.05	0.121	0.355	906 (80%)	0.12	0.071	0.324	284	0.40	0.016	0.286	284	0.40	0.016	0.286
cascade lookup off-by-one	0.615	167	0.05	0.078	0.313	87	0.09	0.032	0.294	167	0.05	0.078	0.312	87	0.09	0.032	0.294
cascade lookup swap	0.196	871 (90%)	0.05	0.113	0.442	493	0.09	0.058	0.360	871 (90%)	0.05	0.113	0.442	493	0.09	0.058	0.360

No change (MRI clean, LLM regeneration): no policy falls back on any stream; audits at floor 0.05 cost 0.016 (MRI) and 0.051 (LLM, whose frozen rule saves about 97% of the samples) of the full budget over doing nothing.

Visible changes near or below the target (MRI drift, cascade noise, LLM model replacement; risks 0.025, 0.05024 and 0.043, with noise slightly above α): the monitor fires on every stream (lower medians 70–83 subjects, at most 145) and the fallback then completes the rest of the stream, at 0.10, 0.08 and 0.67 of the full budget over doing nothing, for two within-target changes and one near-boundary change; alarm-driven escalation stays at the floor and never falls back (0.013, 0.019, 0.048); the joint policy pays the ceiling rate for the rest of the stream (0.057, 0.055, 0.35) and further lowers delivered disagreement. Where the reference is the better output the fallback also lowers the task error (noise $0.175 \to 0.158$, model replacement $0.092 \to 0.071$), which no disagreement certificate sees. *Visible, with excess disagreement* (token cap, risk 0.133): the monitor fallback and the joint policy at floor 0.05 reach a similar cost (0.748 against 0.744), the monitor fallback at the lower delivered disagreement (0.002 against 0.005, lower on every one of the ten streams): a small exchange of cost for risk, not an equivalence; alarm-driven escalation lies between the uniform rates and them (0.681, 0.015), and no audited point at a lower cost reaches the monitor fallback's disagreement.

Missed by the monitor, with excess disagreement (MRI stage-$(k-1)$ image, cascade off-by-one, cascade swap; risks 0.70, 0.62, 0.20; bugs an integrity check would catch given a trusted copy of $\mathcal{D}$): the monitor never fires, so the monitor fallback is the do-nothing point (delivered disagreement 0.35, 0.32, 0.11 over the stream; on the cascade a task error of 0.413 and 0.198 against 0.109 and 0.109 for never stopping); every audited policy moves off it, and the fallback is the cost: alarm-driven escalation at floor 0.05 delivers 0.028, 0.029 and 0.043 at costs 0.847, 0.706 and 0.640, that is 0.16, 0.29 and 0.36 below never stopping, and on the cascade task errors of 0.127 and 0.138. Among the audited points the median escalation points lie below and to the left of the uniform ones where both fall back on every stream (MRI bug, floor 0.02 against $p = 0.05$: cost 0.838 against 0.840 at disagreement 0.037 against 0.039), but paired intervals over the ten streams do not resolve these differences (cost -0.002, 95% bootstrap interval $[-0.008, 0.005]$; disagreement -0.005, $[-0.019, 0.008]$); floor 0.05 against $p = 0.1$ lowers the cost (-0.008 $[-0.010, -0.006]$ on the MRI bug, -0.013 $[-0.019, -0.006]$ on the cascade off-by-one) with no statistically resolved change in disagreement. The two escalations coincide.

Reference failures (MRI low-SNR $2\times$, cascade pixelation): they exceed the disagreement target (the MRI dropout is invisible to the monitor, and all audited policies fall back, with lower-median delays of 20–169 subjects), but on pixelation every evaluated fallback policy raises the delivered task error, from 0.477 for doing nothing to 0.496 (alarm-driven), 0.511 (joint), 0.514 (monitor fallback) and 0.513 (never stop): the certificate reports the disagreement correctly and the decision to complete needs a quality measure. The resulting trade-offs depend on the change. Audit confirmation avoids persistent fallback on the two within-target changes and detects the three bugs missed by the score monitor. Direct monitor fallback is competitive on the token cap. All completion policies require a useful reference, and continuous auditing carries a measurable no-change cost. The never-stopping MRI policy also pays the decision stall, giving normalized cost 1.01. The heuristic scale in Tables 35, 36 and 37 is $\log(1/(\delta w_u))/(p\,\mathrm{kl}(r, \alpha))$, using the last doubling restart before the change and its weight w_u, reported for $r > 1.2\alpha$. It omits pre-change capital loss and adaptive betting effects and is a heuristic comparison, not a delay bound for the betting alarm.

Test subjects and pooled subjects. The stationary MRI monitoring study uses the 59 held-out test subjects; the ranges of its stationary bound over five streams are 0.052–0.098, 0.053–0.082 and 0.055–0.066 at $\bar{p}$=$0.05, 0.1, 0.2$. The anomaly streams of Table 23 pool all 118 calibration and test subjects, on which λ=0.2 has population risk 0.017 (bound 0.052 on $2/118$ against 0.103 on $2/59$), and there the stationary bound after 3000 subjects is 0.057, 0.040 and 0.036 (drift-valid 0.112, 0.068, 0.046), no alarm fires, and net saving is 0.295, 0.278 and 0.247 against 0.311 without audits (Figure 12). Audit randomness moves the bound by a few hundredths across streams; which subjects fail moves it by tenths: the certificate is a statement about the pool being audited, and the number of independent subjects, not the stream length, bounds what it says about a new one. On the 59 test subjects alone the anomalies of Table 23 behave alike: the delivery off-by-one fails 0.678 of the stopped subjects and the LR alarm fires after 148, 72 and 41 subjects at $\bar{p}$=$0.05, 0.1, 0.2$ (betting 245, 107, 86); the reconstructor bypass and the low-SNR dropout after 70, 38 and 15 (156, 76, 59), with

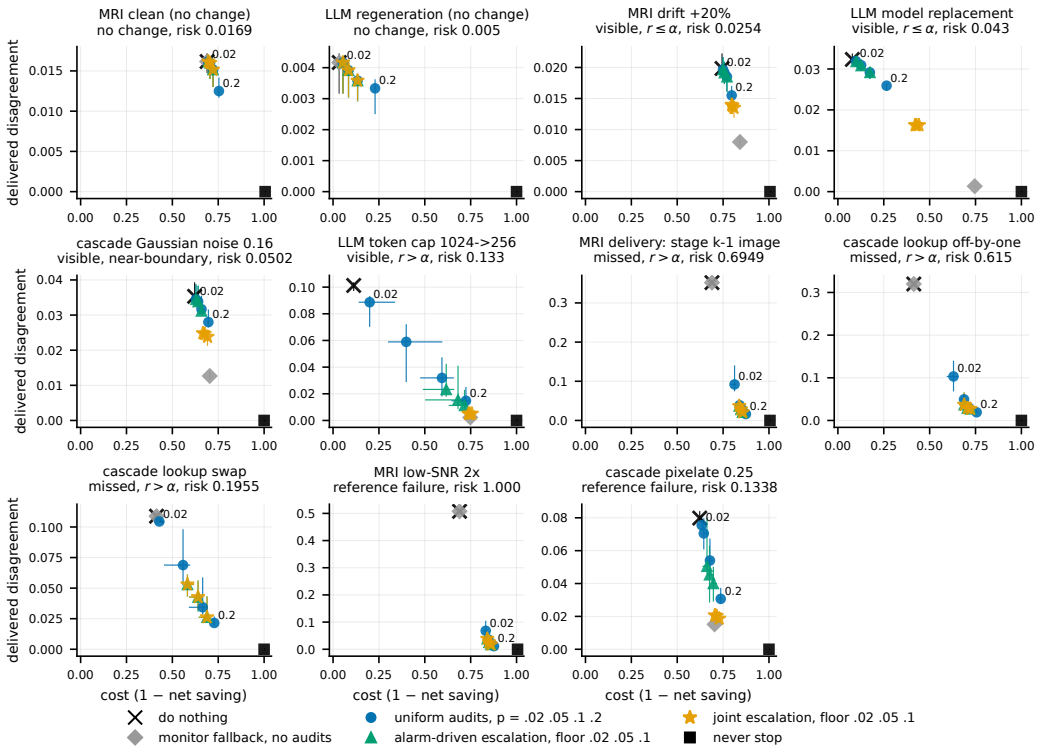

Figure 10: E13 cost versus delivered disagreement: medians and interquartile ranges over 10 streams. Cost includes audits, deferrals and fallback. Titles: condition, post-change disagreement risk r, and whether the mean monitor detects the change (visible/missed). Numeric labels: audit rates.

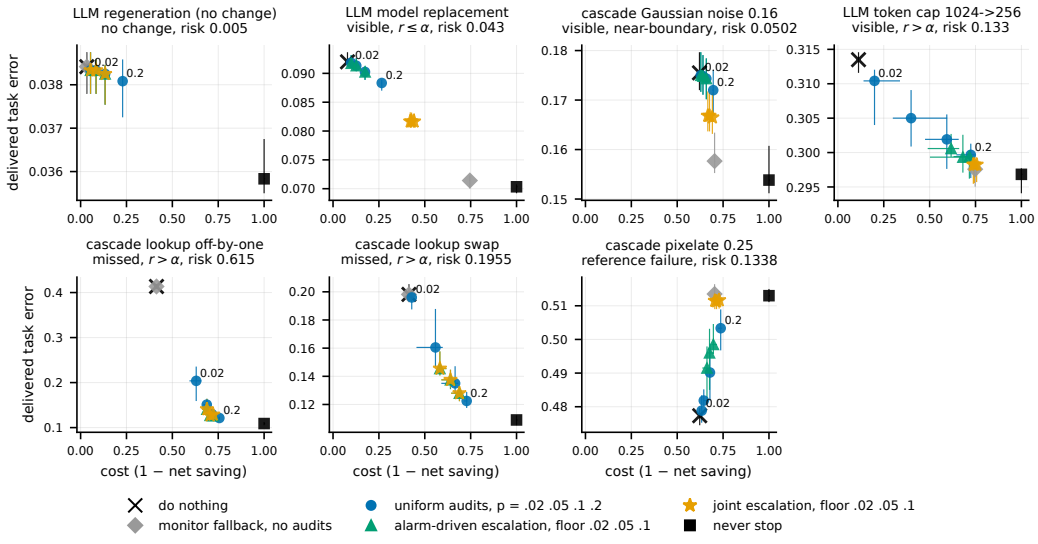

Figure 11: Cost versus delivered task error for E13 conditions with ground truth (GSM8K answers and Imagenette labels). Same policies and summaries as Figure 10.

Table 32: E13: cost / delivered disagreement, medians over 10 streams. Uniform-policy labels give audit rates; adaptive-policy labels give floors, with ceiling 0.5. Conditions follow Figure 10.

	no change		visible, $r \leq \alpha$		visible, near-boundary	visible, $r > \alpha$	missed, $r > \alpha$			reference failure	
policy	MRI clean	LLM regen.	MRI drift $+20\%$	LLM model repl.	casc. noise 0.16	LLM token cap	MRI stage $k-1$	casc. off-by-one	casc. swap	MRI low-SNR $2\times$	casc. pixelate
do nothing	0.69 / 0.016	0.03 / 0.004	0.74 / 0.020	0.08 / 0.032	0.62 / 0.035	0.11 / 0.101	0.69 / 0.352	0.41 / 0.320	0.41 / 0.109	0.69 / 0.508	0.62 / 0.080
monitor fallback	0.69 / 0.016	0.03 / 0.004	0.84 / 0.008	0.75 / 0.001	0.70 / 0.013	0.75 / 0.002	0.69 / 0.352	0.41 / 0.320	0.41 / 0.109	0.69 / 0.508	0.70 / 0.015
uniform 0.02	0.70 / 0.016	0.05 / 0.004	0.75 / 0.020	0.10 / 0.032	0.63 / 0.035	0.20 / 0.089	0.81 / 0.092	0.63 / 0.103	0.43 / 0.105	0.83 / 0.068	0.63 / 0.075
uniform 0.05	0.71 / 0.016	0.08 / 0.004	0.76 / 0.019	0.13 / 0.031	0.64 / 0.034	0.40 / 0.059	0.84 / 0.039	0.69 / 0.050	0.56 / 0.069	0.85 / 0.036	0.64 / 0.071
uniform 0.1	0.72 / 0.015	0.13 / 0.004	0.77 / 0.018	0.17 / 0.029	0.66 / 0.032	0.59 / 0.032	0.85 / 0.028	0.72 / 0.029	0.67 / 0.034	0.86 / 0.022	0.68 / 0.054
uniform 0.2	0.75 / 0.013	0.23 / 0.003	0.80 / 0.015	0.27 / 0.026	0.70 / 0.028	0.72 / 0.015	0.87 / 0.016	0.76 / 0.019	0.73 / 0.021	0.88 / 0.011	0.74 / 0.031
alarm-driven 0.02	0.70 / 0.016	0.05 / 0.004	0.75 / 0.020	0.10 / 0.032	0.63 / 0.035	0.62 / 0.023	0.84 / 0.037	0.69 / 0.037	0.58 / 0.053	0.84 / 0.039	0.66 / 0.051
alarm-driven 0.05	0.71 / 0.016	0.08 / 0.004	0.76 / 0.019	0.13 / 0.031	0.64 / 0.034	0.68 / 0.015	0.85 / 0.028	0.71 / 0.029	0.64 / 0.043	0.85 / 0.026	0.68 / 0.045
alarm-driven 0.1	0.72 / 0.015	0.13 / 0.004	0.77 / 0.018	0.17 / 0.029	0.66 / 0.031	0.72 / 0.011	0.86 / 0.021	0.72 / 0.027	0.69 / 0.026	0.86 / 0.018	0.70 / 0.040
joint 0.02	0.70 / 0.016	0.05 / 0.004	0.80 / 0.014	0.42 / 0.016	0.67 / 0.025	0.74 / 0.005	0.84 / 0.037	0.69 / 0.037	0.58 / 0.053	0.84 / 0.039	0.70 / 0.021
joint 0.05	0.71 / 0.016	0.08 / 0.004	0.80 / 0.014	0.43 / 0.016	0.68 / 0.025	0.74 / 0.005	0.85 / 0.028	0.71 / 0.029	0.64 / 0.043	0.85 / 0.026	0.71 / 0.020
joint 0.1	0.72 / 0.015	0.13 / 0.004	0.81 / 0.013	0.44 / 0.016	0.69 / 0.024	0.76 / 0.005	0.86 / 0.021	0.72 / 0.027	0.69 / 0.026	0.86 / 0.018	0.73 / 0.019
never stop	1.01 / 0.000	1.00 / 0.000	1.01 / 0.000	1.00 / 0.000	1.00 / 0.000	1.00 / 0.000	1.01 / 0.000	1.00 / 0.000	1.00 / 0.000	1.01 / 0.000	1.00 / 0.000

Table 33: E13: fallback fraction and, in parentheses, lower-median post-change delay in subjects. + marks censoring at the observation window. Fallbacks before the change count toward the fraction but not the delay.

	no change		visible, $r \leq \alpha$		visible, near-boundary	visible, $r > \alpha$	missed, $r > \alpha$		reference failure	
policy	MRI clean	LLM regen.	MRI drift +20%	LLM model repl.	casc. noise 0.16	LLM token cap	MRI stage $k-1$	casc. off-by-one	casc. swap	MRI low-SNR 2×	casc. pixelate
do nothing	–	–	–	–	–	–	–	–	–	–	–
monitor fallback	0%	0%	100% (81)	100% (83)	100% (70)	100% (68)	0%	0%	0%	0%	100% (72)
uniform 0.02	0%	0%	0%	0%	0%	70% (3838)	100% (356)	100% (456)	20% (1500+)	100% (169)	20% (1500+)
uniform 0.05	0%	0%	0%	0%	0%	90% (2443)	100% (141)	100% (167)	90% (871)	100% (89)	40% (1500+)
uniform 0.1	0%	0%	0%	0%	0%	100% (1034)	100% (100)	100% (93)	100% (433)	100% (44)	80% (1101)
uniform 0.2	0%	0%	0%	0%	0%	100% (657)	100% (59)	100% (52)	100% (221)	100% (20)	100% (569)
alarm-driven 0.02	0%	0%	0%	0%	0%	90% (1116)	100% (109)	100% (123)	100% (745)	100% (94)	60% (952)
alarm-driven 0.05	0%	0%	0%	0%	0%	100% (639)	100% (88)	100% (87)	100% (493)	100% (55)	80% (906)
alarm-driven 0.1	0%	0%	0%	0%	0%	100% (530)	100% (67)	100% (73)	100% (287)	100% (31)	80% (741)
joint 0.02	0%	0%	0%	0%	0%	100% (324)	100% (109)	100% (123)	100% (744)	100% (94)	100% (284)
joint 0.05	0%	0%	0%	0%	0%	100% (324)	100% (88)	100% (87)	100% (493)	100% (55)	100% (284)
joint 0.1	0%	0%	0%	0%	0%	100% (325)	100% (67)	100% (76)	100% (287)	100% (31)	100% (232)
never stop	–	–	–	–	–	–	–	–	–	–	–

the same guardrail rates as on the full pool; the 20% drift has risk 0.051 on this pool and the low-SNR population shift 0.068, and neither raises an LR alarm within 1500 subjects (the betting alarm fires on one stream in ten at $\bar{p}$=0.2 for the population shift): small violations also remain undetected within the tested window, as in E8.

Across repeated splits (E11). We evaluate nine additional splits of the same 183 subjects (seeds 20260901–20260909), each with its own denoiser, tolerance and trajectory cache. Training, calibration and test groups are disjoint within each split; subjects recur across splits. Table 38 repeats E1 and the two delivery-path bugs on each, with the threshold selected on that split's calibration subjects. The certified threshold is 0.2 on six splits, 0.1 on two and 1.0 on one; its population risk is 0–0.017 (median 0.008), no alarm fires on any stationary stream, and after 3000 subjects the stationary bound at $\bar{p}$=0.1 is 0.014–0.049 (drift-valid 0.040–0.075). The off-by-one delivery bug fails 0.48–0.73 of the stopped subjects, is invisible to the label-free monitor on every split, and is caught by the LR alarm after 57–115 subjects at $\bar{p}$=0.1 (90–233 at 0.05, 22–73 at 0.2; betting 114–164 at 0.1) on every stream, the guardrail bringing the delivered failure rate from 0.43–0.65 to 0.027–0.043; the reconstructor bypass reproduces the main split's numbers on every split (38–39 subjects, 0.87–0.90 → 0.028–0.033). The spread across splits comes from which subjects fail at the split's tolerance, not from the audits. In Table 41, the fixed-level disagreement alarm triggers

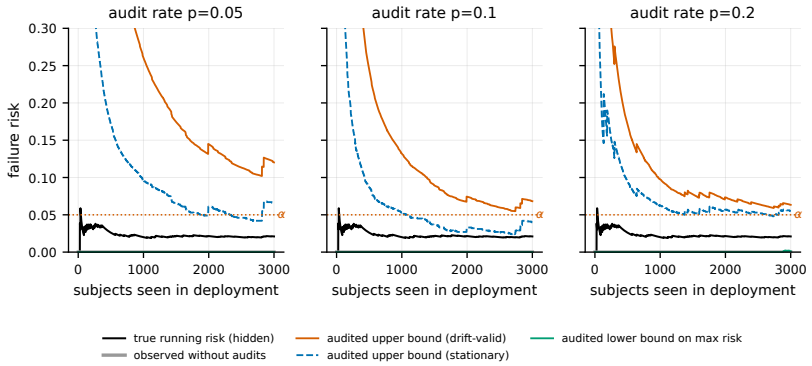

Figure 12: Stationary M4Raw replay (3000 subjects, pooled calibration and test cohort). Black: running risk; grey: observed loss without audits. Dashed/solid upper bounds: stationary betting/drift-valid EB. Green: lower bound on maximum risk.

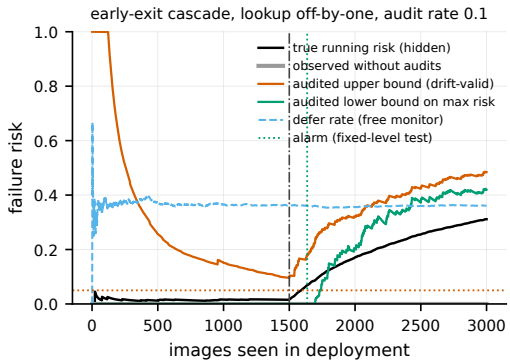

Figure 13: Cascade label-mapping failure on one stream. Dashed: failure begins; dotted: audited betting alarm. Confidence and exit rates remain unchanged. Cross-stream detection delays appear in Tables 39 and 37.

persistent full acquisition. The certification veto additionally returns to $\lambda^{(0)}$ whenever the drift-valid upper bound exceeds α. For comparison, the closed loop pays 114 audits at $\bar{p} = 0.1$ and 268 at $\bar{p} = 0.2$ on these same streams.

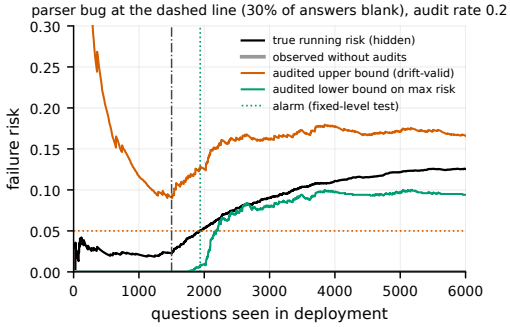

Figure 14: LLM parser failure on one stream: 30% of answers are blanked after the dashed line. Disagreement rises and the audited alarm fires. Bounds versus audit count appear in Figure 16.

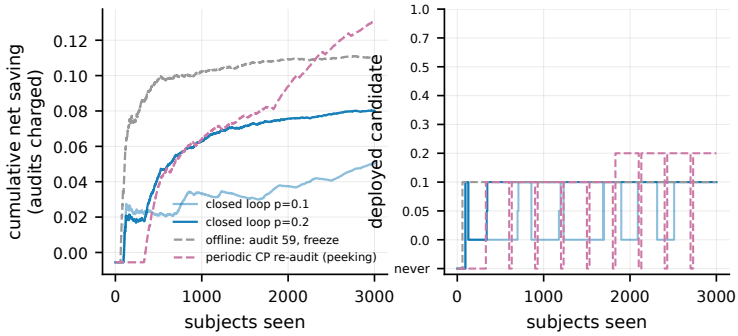

Figure 15: M4Raw deployment from full acquisition (tolerance quantile 0.5). Left: cumulative net saving, including audits. Right: deployed candidate. Offline calibration uses 59 subjects; the closed loop continues updating.

LLM domain. Each question is answered with $N{=}64$ independent samples from Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024) (chat template, system prompt "Please reason step by step, and put your final answer within \boxed{}", temperature 0.8, top-p 0.95, at most 1024 new tokens, fixed seed; vLLM). The answer is the last \boxed{} expression, normalised (whitespace, currency and percent signs, \text, \dfrac, \left/\right, degree marks, thousands separators; GSM8K answers parsed to numbers). If no box is present, the extractor uses the last number; if neither is usable, the sample votes for the empty string. The full-budget output is the plurality answer at 64 with ties broken by first occurrence; stopping at k delivers the plurality at k; a question fails if the two differ. Checkpoints are $k = 2, \ldots, 63$. The stopping score at k draws 64 suffixes of $64 - k$ votes from a Dirichlet-multinomial posterior over the observed answer categories plus one unseen pseudo-category (prior mass $\frac{1}{2}$ each) and reports the fraction of draws in which the final plurality differs from the current one. The offline threshold is selected on the first 400 GSM8K test questions (fully sampled) and deployed on the remaining 919, resampled into streams of 3000 (stationary) or 6000 questions (shifts at 1500). The task shift cycles the 262 MATH-500 questions of level 4–5 into a 919-entry pool (so the first 133 appear four times and the rest three) and resamples it; the serving bug overwrites the second sample with a copy of the first; the parser bug blanks each answer with probability 0.3; the misrouting bug replaces the first two samples by two copies of the previous request's first answer, creating a forced-unanimity stress test. Costs count samples (64 per full budget); an audit draws the remaining samples. Single-sample accuracy is 0.855 on GSM8K and 0.705 on MATH-500; the 64-sample plurality reaches 0.903 and 0.768. The plurality at k agrees with the plurality at 64 for 90/94/96/98/99% of GSM8K questions at $k = 2/4/8/16/32$ and for 80/85/88/91/95% of MATH-500 questions, which measures how often additional samples change the proposed answer. Samples truncated at the token cap (0.4% on GSM8K, 7.3% on MATH-500) remain in the pool and use the same extraction rule. A last-number fallback can turn an intermediate calculation into a vote.

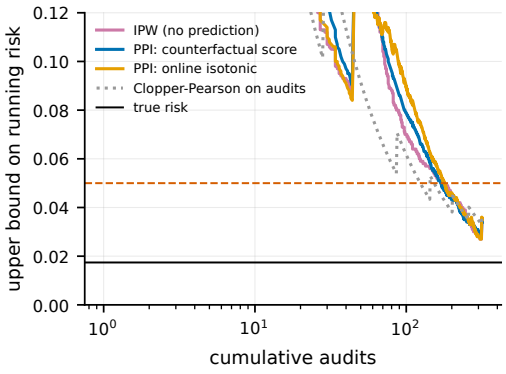

Figure 16: LLM domain: certificate as a function of cumulative audits on the GSM8K deployment stream, $\bar{p}$=0.1, three predictors.

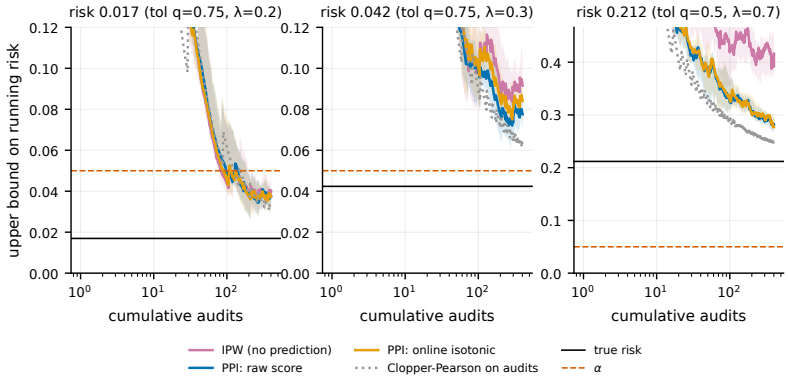

Figure 17: Risk upper bounds versus audit count for three predictors and the exact binomial baseline. Medians and interquartile ranges over 10 M4Raw streams; $\bar{p} = 0.1$.

Active audits. Allocating the audit rate as $p_i \propto \sqrt{f_i(1-f_i)}$ (normalised to $\bar{p}$, floor 0.02) made no measurable difference to the certificate relative to uniform audits at the same average rate on the frequent-failure candidates of the stricter tolerance (10 streams of 4000): at $\bar{p}$=0.1 the median bound after 4000 subjects was 0.269 versus 0.275 (risk 0.203) and 0.273 versus 0.280 (risk 0.212); at $\bar{p}$=0.2, 0.257 versus 0.253 and 0.266 versus 0.262. On these candidates, residual variance varies little with the raw score, which limits the gain from this score-based allocation. Score-dependent audits are therefore left for a calibrated predictor; the certificate and the betting alarm hold for either choice, whereas the LR alarm needs a rate fixed before the subject is seen and is disabled under score-dependent audits.

Early-exit cascade. Stages are torchvision's ResNet-18 (IMAGENET1K_V1), ResNet-50 (IMAGENET1K_V2) and ConvNeXt-Base (IMAGENET1K_V1), run on the Imagenette-320 validation split (ten ImageNet classes, 3925 images drawn from ImageNet-val, so none were used in training) after resizing to 256 and centre-cropping to 224. Checkpoints are stages 1 and 2; the rule exits at the first stage whose $1-$max-softmax is at or below the threshold; the unit fails if the exited label differs from ConvNeXt-Base's label. Costs are cumulative GFLOPs (1.8, 5.9, 21.3); an audit runs the remaining stages. Corruptions are applied to the cropped image before normalisation: additive Gaussian noise ($\sigma = 0.08, 0.16$), Gaussian blur (radius $2, 4$), JPEG re-encoding (quality $10, 5$), contrast scaling (0.3), brightness offset ($+0.3$) and pixelation (0.25). The cache is stored in class order, so a fixed random permutation (seed 20260911) is applied before the split; the offline threshold is selected on the first 800 permuted images (all ten classes represented) and streams of 3000 resample the remaining 3125, with the corruption applied to the same image indices from 1500

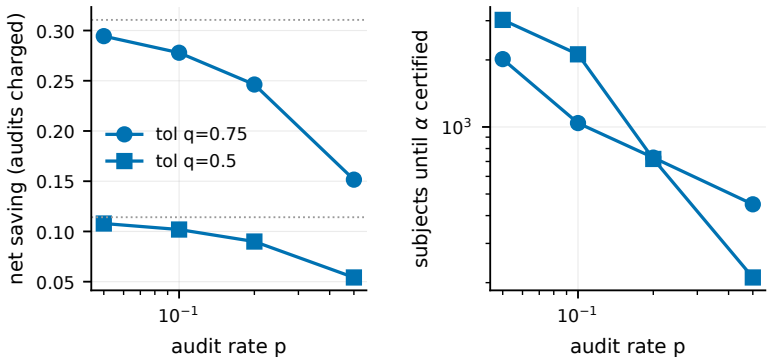

Figure 18: M4Raw audit-rate trade-off at two tolerances. Left: net saving, including audits; dotted lines omit audit cost. Right: subjects required to certify risk at most α.

on. The two integration bugs act on the recorded outputs: an off-by-one bug maps each stage-1 label in the ten-class set to the next class of the set (an off-by-one in the task's lookup table); a two-class swap exchanges the two most frequent stage-1 classes (tench and golf ball, whose stage-1 frequencies differ by 0.15 percentage points: 368 against 362 of 3925 images). Both leave stage-1 confidence, exit rate and label support unchanged; the swap also leaves the label histogram unchanged up to a total variation of 0.001. We report final-stage accuracy alongside disagreement. When the reference collapses, a valid disagreement certificate alone cannot justify fallback.

Streams and seeds. MRI deployment streams of length T are drawn with replacement from the 118 calibration and test subjects (seeded); shift streams switch at t_0 to the corrupted trajectories of the same subject indices. All audit coins, predictor updates and selections are seeded; medians over seeds are reported with interquartile bands in figures.

Predictors. "IPW" uses $f \equiv 0$; "raw score" uses the stopping score itself (a frequency in $[0, 1]$); "online isotonic" refits a pool-adjacent-violators calibration of score to audited failure every five audited subjects (quantile bins, Laplace smoothing) and falls back to the raw score until 20 audited subjects are available. Subject-level predictions combine contrasts as $1 - \prod_c (1 - f_c)$.

Label-free change monitor. The baseline watches per-subject statistics that need no audit: the mean proposed-stage fraction $d_i / R_{\max}$ over the subject's units and the stopping score at the first checkpoint (clipped to $[0, 1]$); for the cascade a separate monitor watches the delivered label's one-hot indicators (the label histogram). Each statistic is monitored as follows, with a total budget δ per monitor split equally over its statistics. The first 300 subjects give a two-sided betting confidence interval $[\ell_0, h_0]$ for their mean at level $\delta'/2$ (δ' the statistic's budget); afterwards two fixed-level restarted betting tests of the form of Corollary 1(i), with the disagreement alarm's restart schedule counted from the end of the burn-in and the same bet sizing, test excess mean at threshold $h_0 + \varepsilon$ on x and at threshold $1 - \ell_0 + \varepsilon$ on $1 - x$, each at level $\delta'/4$. Under the null that the burn-in conditional means are constant and every later conditional mean lies within ε of that constant, the probability of any alarm over the deployment is at most δ'. We report first-ever alarms: an alarm before the injected shift is a false alarm, not a delay. The dead zone is $\varepsilon = 0.02$ in the tables (0.05 as a sensitivity in the reports); on 50 unshifted M4Raw streams the monitor raised 0 alarms (one-sided 95% upper bound 0.058 on the rate). Its null concerns conditional means. Changes that preserve those means or remain within ε do not violate that null. If the entire prefix law is unchanged, Proposition 1 additionally rules out distinguishing the two laws from prefix observations alone.

Confidence sequences. Restart schedules below use zero-based processed-subject counts: a component started at count t first bets on subject $t + 1$ in the theory's one-based indexing. Betting: the upper-CS factors are $1 + \eta_i^{\mathrm{up}}(\theta)(\theta - Z_i)$, with variance-adaptive fractions (Waudby-Smith & Ramdas, 2024) $\eta_i^{\mathrm{up}}(\theta) = \mathrm{clip}\big((\theta - \hat{\mu}_{i-1})/(\hat{\sigma}_{i-1}^2 + (\theta - \hat{\mu}_{i-1})^2),\, 0,\, \kappa/(h_i - \theta)\big)$ with $\kappa = 0.75$, predictable running moments with a half pseudo-count, a grid of 1001 candidate risks on $[0, 1]$,

bound rounded outward by one grid point; $(\ell_i, h_i) = (0, 1)$ while the deployed candidate is $\lambda^{(0)}$ and $(1 - 1/\bar{p}_i,\ 1/\bar{p}_i)$ otherwise, $\bar{p}_i$ the scheduled audit rate of the step (the constant rate under uniform audits). Betting disagreement alarm: capital processes restarted at $t \in \{0, 50, 100, 200, 400, \dots\}$ with weights $6/(\pi^2 k^2)$, bets sized from exponentially weighted moments (half-life 100 subjects) and mixed over multipliers $\{1, 3, 9\}$ of the corresponding lower-bound fraction, all clipped at $\kappa/(\theta - \ell_i)$. EB: $\mathcal{V} = \{0.9, 0.75, 2^{-1}, \dots, 2^{-12}\}$ ($|\mathcal{V}| = 14$), centre = running mean with a half pseudo-count, global range $[1 - 1/p_{\min}, 1/p_{\min}]$ with $p_{\min}$ the scheduled floor (the rate under uniform audits, the floor under escalation, the global lower propensity bound under active audits).

LR grid and restarts. The alternatives are $\{0.1, 0.2, 0.4, 0.8, 0.95\}$ at $\alpha = 0.05$, with uniform weights. For $r \geq r_g$, the derivative of $I_g(r)/\mathrm{kl}(r, \alpha)$ has numerator $I_g'(r)\mathrm{kl}(r, \alpha) - I_g(r)\partial_r \mathrm{kl}(r, \alpha)$. This numerator vanishes at $r = r_g$ and decreases thereafter: its derivative is $-I_g(r)\partial_r^2 \mathrm{kl}(r, \alpha) < 0$. The ratio therefore decreases on each grid interval. Its infima are 0.682, 0.812, 0.824, 0.965, 0.983 on the five intervals from 0.1 to 1, using the largest grid point at or below r. Thus it is at least 0.68 for $r \geq 0.1$, with the worst value approached as $r \uparrow 0.2$. The smallest alternative's positive-drift threshold is approximately 0.0724.

Restarts use the doubling schedule $\{0, 50, 100, 200, \dots\}$ with weights $3/(\pi^2 u^2)$ and on the linear schedule $\{0, 50, 100, 150, \dots\}$ with weights $1/(2u(u+1))$, dormant components counted at capital one; the delay bound in the tables uses the first linear restart at or after the change (1500, 1550, 1550 for changes at 1500, 1501, 1533; $u = 31, 32, 32$) with its own weight and adds the waiting time (0, 49, 17); bounds are rounded up. Theorem 1(ii) gives a sufficient rate for an expected delay of at most $\tau^\star$ from that restart, on shifts to risk at least $r_{\min} \geq 2\alpha$: $p \geq [\log(G/(\delta w')) + \log(r_g/\alpha)]/(\tau^\star I_g(r_{\min}))$. For the linear restart after 1500 subjects, $\delta = 0.05$ and $\tau^\star = 500$, this gives $p \geq 0.195$ at $r = 0.2$ or $p \geq 0.036$ at $r = 0.5$. At $r = 0.1$ the sufficient rate exceeds one, so the bound cannot certify this delay even with full observation. Escalation: floor $p_{\mathrm{floor}} \in \{0.02, 0.05\}$, $p_{\max} = 0.5$, $e_{i-1} = \max(\log^+ M_{i-1}(\alpha), \log^+ M_{i-1}^{\mathrm{LR}})/\log(1/\delta_a)$ clipped to $[0, 1]$, where $\log^+ x = \max\{0, \log x\}$ and $\delta_a = \delta/2$ when both alarms are active. Single-alarm policies use only that alarm's capital and $\delta_a = \delta$. We use $\phi(e) = \sqrt{e}$.

Table 34: E13: cost / delivered task error (medians over 10 streams) on the conditions with ground truth.

policy	no change LLM regen.	visible, $r \leq \alpha$ LLM model repl.	visible, near-boundary casc. noise 0.16	visible, $r > \alpha$ LLM token cap	missed, $r > \alpha$ casc. off-by-one	 casc. swap	reference failure casc. pixelate
do nothing	0.03 / 0.038	0.08 / 0.092	0.62 / 0.175	0.11 / 0.314	0.41 / 0.413	0.41 / 0.198	0.62 / 0.477
monitor fallback	0.03 / 0.038	0.75 / 0.071	0.70 / 0.158	0.75 / 0.298	0.41 / 0.413	0.41 / 0.198	0.70 / 0.514
uniform 0.02	0.05 / 0.038	0.10 / 0.092	0.63 / 0.175	0.20 / 0.310	0.63 / 0.204	0.43 / 0.196	0.63 / 0.479
uniform 0.05	0.08 / 0.038	0.13 / 0.091	0.64 / 0.175	0.40 / 0.305	0.69 / 0.151	0.56 / 0.161	0.64 / 0.482
uniform 0.1	0.13 / 0.038	0.17 / 0.090	0.66 / 0.174	0.59 / 0.302	0.72 / 0.130	0.67 / 0.135	0.68 / 0.490
uniform 0.2	0.23 / 0.038	0.27 / 0.088	0.70 / 0.172	0.72 / 0.300	0.76 / 0.121	0.73 / 0.122	0.74 / 0.503
alarm-driven 0.02	0.05 / 0.038	0.10 / 0.092	0.63 / 0.175	0.62 / 0.301	0.69 / 0.141	0.58 / 0.145	0.66 / 0.491
alarm-driven 0.05	0.08 / 0.038	0.13 / 0.091	0.64 / 0.175	0.68 / 0.299	0.71 / 0.127	0.64 / 0.138	0.68 / 0.496
alarm-driven 0.1	0.13 / 0.038	0.17 / 0.090	0.66 / 0.174	0.72 / 0.299	0.72 / 0.126	0.69 / 0.128	0.70 / 0.498
joint 0.02	0.05 / 0.038	0.42 / 0.082	0.67 / 0.167	0.74 / 0.298	0.69 / 0.141	0.58 / 0.145	0.70 / 0.512
joint 0.05	0.08 / 0.038	0.43 / 0.082	0.68 / 0.167	0.74 / 0.298	0.71 / 0.127	0.64 / 0.138	0.71 / 0.511
joint 0.1	0.13 / 0.038	0.44 / 0.082	0.69 / 0.167	0.76 / 0.298	0.72 / 0.129	0.69 / 0.128	0.73 / 0.512
never stop	1.00 / 0.036	1.00 / 0.070	1.00 / 0.154	1.00 / 0.297	1.00 / 0.109	1.00 / 0.109	1.00 / 0.513

Table 35: MRI companion to Table 23: LR delay bound, betting delay and heuristic scale, escalated audits, and net saving without/with fallback. "@" separates delay from paid audit rates before/after change; parentheses give detection fractions below one.

(a) Risk and LR delay bound

Anomaly	Risk	Defer rate	LR bound: $p = 0.05$ / 0.1 / 0.2
clean control	0.017	0.00	– / – / –
repetition order permuted	0.000	0.02	– / – / –
intensity drift $+20\%$	0.025	0.16	– / – / –
intensity drift $+40\%$	0.000	1.00	– / – / –
motion spike 4 px	0.000	1.00	– / – / –
motion spike 8 px	0.000	1.00	– / – / –
ghosting 20%	0.000	1.00	– / – / –
ghosting 40%	0.000	1.00	– / – / –
last rep. low-SNR $1\times$	1.000	0.00	103 / 52 / 26
last rep. low-SNR $2\times$	1.000	0.00	103 / 52 / 26
delivery: stage $k-1$ image	0.695	0.00	207 / 104 / 52
delivery: no reconstructor	1.000	0.00	103 / 52 / 26
population: 40 lowest-SNR	0.051	0.00	– / – / –

(b) Betting delay (detection fraction) / heuristic scale

Anomaly	$p = 0.05$	$p = 0.1$	$p = 0.2$
clean control	–	–	–
repetition order permuted	–	–	–
intensity drift $+20\%$	–	–	–
intensity drift $+40\%$	–	–	–
motion spike 4 px	–	–	–
motion spike 8 px	–	–	–
ghosting 20%	–	–	–
ghosting 40%	–	–	–
last rep. low-SNR $1\times$	162 / 47	86 / 24	66 / 12
last rep. low-SNR $2\times$	162 / 47	86 / 24	66 / 12
delivery: stage $k-1$ image	245 / 95	128 / 48	86 / 24
delivery: no reconstructor	162 / 47	86 / 24	66 / 12
population: 40 lowest-SNR	–	–	–

(c) Escalated audits and net saving

Anomaly	LR floor 0.05	Betting floor 0.05	Betting floor 0.02	Net saving ($p = 0.1$) frozen → fallback
clean control	– @ 0.05→0.05	– @ 0.05→0.05	– @ 0.02→0.02	0.278 → 0.278
repetition order permuted	– @ 0.05→0.05	– @ 0.05→0.05	– @ 0.02→0.02	0.264 → 0.264
intensity drift $+20\%$	– @ 0.05→0.04	– @ 0.05→0.04	– @ 0.02→0.02	0.229 → 0.229
intensity drift $+40\%$	– @ 0.05→0.00	– @ 0.05→0.00	– @ 0.02→0.00	0.137 → 0.137
motion spike 4 px	– @ 0.05→0.00	– @ 0.05→0.00	– @ 0.02→0.00	0.137 → 0.137
motion spike 8 px	– @ 0.05→0.00	– @ 0.05→0.00	– @ 0.02→0.00	0.137 → 0.137
ghosting 20%	– @ 0.05→0.00	– @ 0.05→0.00	– @ 0.02→0.00	0.137 → 0.137
ghosting 40%	– @ 0.05→0.00	– @ 0.05→0.00	– @ 0.02→0.00	0.137 → 0.137
last rep. low-SNR $1\times$	54 @ 0.05→0.49	63 @ 0.05→0.49	109 @ 0.02→0.48	0.278 → 0.141
last rep. low-SNR $2\times$	54 @ 0.05→0.49	63 @ 0.05→0.49	109 @ 0.02→0.48	0.278 → 0.141
delivery: stage $k-1$ image	88 @ 0.05→0.48	91 @ 0.05→0.48	141 @ 0.02→0.47	0.278 → 0.146
delivery: no reconstructor	54 @ 0.05→0.49	63 @ 0.05→0.49	109 @ 0.02→0.48	0.278 → 0.141
population: 40 lowest-SNR	– @ 0.05→0.05	– @ 0.05→0.05	– @ 0.02→0.02	0.269 → 0.269

Table 36: Betting disagreement alarm on the LLM shifts (companion to Table 44).

(a) Risk and betting delay (detection fraction) / heuristic scale

Condition	Risk	$p = 0.05$	$p = 0.1$	$p = 0.2$
task shift: MATH-500 level ≥ 4	0.082	4500+ (30%) / 15894	3547 (60%) / 7947	2850 (70%) / 3974
serving bug: first sample duplicated	0.100	4301 (50%) / 6826	2717 (70%) / 3413	1394 / 1706
parser bug: 30% answers blank	0.161	1282 / 1683	1050 / 842	528 / 421
serving bug: first two responses misrouted	0.986	162 / 49	86 / 25	59 / 12

(b) Escalated betting: delay @ paid rate before/after change

Condition	Floor 0.05	Floor 0.02	Net saving ($p = 0.1$) frozen → fallback
task shift: MATH-500 level ≥ 4	3765 (60%) @ 0.05→0.11	3093 (60%) @ 0.02→0.17	0.782 → 0.775
serving bug: first sample duplicated	1825 (60%) @ 0.05→0.24	4249 (50%) @ 0.02→0.07	0.863 → 0.676
parser bug: 30% answers blank	518 @ 0.05→0.45	666 @ 0.02→0.42	0.827 → 0.389
serving bug: first two responses misrouted	57 @ 0.05→0.50	109 @ 0.02→0.49	0.863 → 0.219

Table 37: Betting disagreement alarm on the cascade corruptions and bugs (companion to Table 39).

(a) Risk and betting delay (detection fraction) / heuristic scale

Condition	Risk	$p = 0.05$	$p = 0.1$	$p = 0.2$
gauss noise 0.08	0.028	–	–	–
gauss noise 0.16	0.050	–	–	–
blur 2.0	0.029	–	–	–
blur 4.0	0.023	–	–	–
jpeg 10	0.027	–	–	–
jpeg 5	0.033	–	–	–
contrast 0.3	0.032	–	–	–
brightness 0.3	0.027	–	–	–
pixelate 0.25	0.134	1500+ (20%) / 2739	1007 / 1370	753 / 685
lookup off-by-one	0.615	233 / 118	147 / 59	95 / 30
two-class swap	0.196	976 (70%) / 1065	675 / 533	389 / 266

(b) Escalated betting: delay @ paid rate before/after change

Condition	Floor 0.05	Floor 0.02	Net saving ($p = 0.1$) frozen → fallback
gauss noise 0.08	– @ 0.03→0.02	– @ 0.01→0.01	0.436 → 0.436
gauss noise 0.16	– @ 0.03→0.01	– @ 0.01→0.00	0.342 → 0.342
blur 2.0	– @ 0.03→0.02	– @ 0.01→0.01	0.407 → 0.407
blur 4.0	– @ 0.03→0.01	– @ 0.01→0.00	0.321 → 0.321
jpeg 10	– @ 0.03→0.02	– @ 0.01→0.01	0.416 → 0.416
jpeg 5	– @ 0.03→0.01	– @ 0.01→0.00	0.321 → 0.321
contrast 0.3	– @ 0.03→0.02	– @ 0.01→0.01	0.443 → 0.443
brightness 0.3	– @ 0.03→0.03	– @ 0.01→0.01	0.491 → 0.491
pixelate 0.25	820 (90%) @ 0.03→0.05	1500+ (10%) @ 0.01→0.00	0.337 → 0.316
lookup off-by-one	79 @ 0.03→0.31	162 @ 0.01→0.30	0.527 → 0.284
two-class swap	600 @ 0.03→0.20	896 (60%) @ 0.01→0.12	0.527 → 0.377

Table 38: MRI across nine additional splits (E11): 5 stationary and 10 shift streams per split, $\bar{p} = 0.1$. Bounds use 3000 subjects; delays are lower medians. Delivered disagreement is shown without/with fallback. Last row: medians across splits.

split seed	λ	risk	bound at 3000, $\bar{p}$=0.1		delivery bug: stage $k-1$ image ($\bar{p}$=0.1)				no reconstructor	
			betting	EB	rule risk	LR	betting	delivered: frozen → guardrail	rule risk	LR
20260901	0.2	0.008	0.025	0.053	0.703	70	122	0.631 → 0.035	1.000	38
20260902	0.2	0.008	0.025	0.052	0.712	59	124	0.627 → 0.030	1.000	38
20260903	0.2	0.000	0.014	0.041	0.653	84	129	0.579 → 0.034	1.000	38
20260904	0.1	0.000	0.014	0.040	0.475	100	164	0.428 → 0.041	0.983	39
20260905	0.1	0.000	0.014	0.040	0.695	68	114	0.625 → 0.038	0.975	38
20260906	0.2	0.017	0.049	0.075	0.695	70	124	0.622 → 0.034	1.000	38
20260907	0.2	0.008	0.024	0.052	0.593	115	148	0.536 → 0.043	1.000	38
20260908	1.0	0.008	0.028	0.056	0.729	57	114	0.650 → 0.027	1.000	38
20260909	0.2	0.000	0.016	0.043	0.653	75	148	0.589 → 0.036	1.000	38
median	–	0.008	0.024	0.052	0.695	70	124	0.622 → 0.035	1.000	38

Table 39: Cascade corruptions and label-mapping bugs at image 1500 of 3000 (10 streams). Column definitions follow Table 23. Integrity checks report expected images until a label/logit mismatch; betting results appear in Table 37.

(a) Task accuracy, rule risk and checks without audits

Condition	Reference accuracy	Risk	Defer rate	Score monitor	Label monitor	Integrity check
gauss noise 0.08	0.90	0.028	0.57	346	–	–
gauss noise 0.16	0.80	0.050	0.80	65	–	–
blur 2.0	0.78	0.029	0.64	133	–	–
blur 4.0	0.41	0.023	0.86	58	1500+ (20%)	–
jpeg 10	0.76	0.027	0.62	186	–	–
jpeg 5	0.57	0.033	0.85	55	–	–
contrast 0.3	0.89	0.032	0.55	425	–	–
brightness 0.3	0.88	0.027	0.43	–	–	–
pixelate 0.25	0.08	0.134	0.82	69	510	–
lookup off-by-one	0.89	0.615	0.35	–	–	2
two-class swap	0.89	0.196	0.35	–	–	6

(b) LR delay (detection fraction) / bound

Condition	$p = 0.05$	$p = 0.1$	$p = 0.2$
gauss noise 0.08	–	–	–
gauss noise 0.16	–	–	–
blur 2.0	–	–	–
blur 4.0	–	–	–
jpeg 10	–	–	–
jpeg 5	–	–	–
contrast 0.3	–	–	–
brightness 0.3	–	–	–
pixelate 0.25	1500+ (10%) / 5620	1288 (50%) / 2810	817 / 1405
lookup off-by-one	128 / 260	77 / 130	43 / 65
two-class swap	1072 (50%) / 2046	576 / 1023	284 / 512

(c) Escalated LR and delivered disagreement

Condition	Floor 0.05: delay @ paid rate before/after	Delivered disagreement ($p = 0.1$) frozen → fallback
gauss noise 0.08	– @ 0.03→0.02	0.025 → 0.025
gauss noise 0.16	– @ 0.03→0.01	0.043 → 0.043
blur 2.0	– @ 0.03→0.02	0.027 → 0.027
blur 4.0	– @ 0.03→0.01	0.021 → 0.021
jpeg 10	– @ 0.03→0.02	0.024 → 0.024
jpeg 5	– @ 0.03→0.01	0.026 → 0.026
contrast 0.3	– @ 0.03→0.02	0.029 → 0.029
brightness 0.3	– @ 0.03→0.03	0.024 → 0.024
pixelate 0.25	912 (90%) @ 0.03→0.05	0.118 → 0.091
lookup off-by-one	75 @ 0.03→0.31	0.545 → 0.040
two-class swap	662 @ 0.03→0.20	0.177 → 0.076

Table 40: Early-exit cascade, stationary Imagenette stream (3000 images, 10 streams): median audits until α is certified and FLOPs saving with audits charged.

$\bar{p}$	audits to certify α, IPW / score: median over certifying streams (fraction certifying)		net saving
0.05	62 (20%)	88 (30%)	0.557
0.1	95 (40%)	54 (30%)	0.525
0.2	154 (70%)	155 (90%)	0.466

Table 41: Closed-loop versus offline calibration (20 streams of 3000). The certification veto requires the deployed-risk EB bound to stay below α. Batch k of the δ-spending policy receives $6\delta/(\pi^2 k^2)$.

method	Pr[unsafe deployed]	median net saving	first deployment
closed loop, $\bar{p} = 0.1$	0.000	0.059	94
closed loop, $\bar{p} = 0.1$ + disagreement alarm	0.000	0.059	94
closed loop, $\bar{p} = 0.1$ + certification veto	0.000	0.015	2381
closed loop, $\bar{p} = 0.2$	0.000	0.068	94
closed loop, $\bar{p} = 0.2$ + disagreement alarm	0.000	0.068	94
closed loop, $\bar{p} = 0.2$ + certification veto	0.000	0.051	1130
offline: audit 59, freeze	0.000	0.025	–
periodic batch re-audit, fixed-n CP (peeking)	0.000	0.066	–
periodic batch re-audit, δ-spending CP (valid)	0.000	0.011	–

Table 42: Offline calibration budget on E3 streams (tolerance quantile 0.5; 20 streams of 3000). The first n subjects are fully acquired; fixed-sequence Clopper–Pearson selection then freezes the rule. Net saving includes all n acquisitions.

offline audits n	Pr[unsafe deployed]	median net saving	certified λ (over streams)
59	0.000	0.025	0, 0.1
100	0.000	0.025	0, 0.1, 0.2
150	0.000	0.064	0, 0.1, 0.2
200	0.000	0.104	0, 0.1, 0.2
300	0.000	0.102	0, 0.1, 0.2
400	0.000	0.098	0.1, 0.2
600	0.000	0.090	0.1, 0.2
900	0.000	0.078	0.1, 0.2
closed loop, $\bar{p} = 0.1$ (E3)	0.000	0.059	adaptive
closed loop, $\bar{p} = 0.2$ (E3)	0.000	0.068	adaptive

Table 43: Near-threshold synthetic cohort (never-stop plus candidates of risk 0, 0.019, 0.039, 0.061, 0.091; α=δ=0.05; 200 streams of 3000).

method	Pr[unsafe ever deployed]	fraction of time unsafe	median net saving
closed loop (anytime CS), $\bar{p} = 0.2$	0.005	0.000	0.035
+ disagreement alarm	0.005	0.000	0.035
+ certification veto (EB)	0.000	0.000	0.021
fixed-n CP on an iid audit subsample ($\bar{p} = 0.2$), re-checked every step (peeking)	0.070	0.000	0.036
fixed-n CP pooling all observed losses, re-checked every step (selective labels + peeking)	0.660	0.287	0.078
offline: audit 59, freeze	0.030	0.000	0.023
periodic batch re-audit, fixed-n CP (peeking)	0.045	0.000	0.034
periodic batch re-audit, δ-spending CP	0.000	0.000	0.007

Table 44: LLM deployment shifts at question 1500 of a 6000-question stream (10 streams). Columns as in Table 23; the betting alarm is in Table 36.

(a) Rule risk and label-free monitoring

Condition	Risk before → after	Samples/question after change	Monitor delay
task shift: MATH-500 level ≥ 4	0.016→0.082	9.6	372
serving bug: first sample duplicated	0.016→0.100	2.0	498
parser bug: 30% answers blank	0.016→0.161	5.6	192
serving bug: first two responses misrouted	0.016→0.986	2.0	557

(b) LR delay (detection fraction) / bound

Condition	$p = 0.05$	$p = 0.1$	$p = 0.2$
task shift: MATH-500 level ≥ 4	4500+ (10%) / 37294	4500+ (40%) / 18647	3261 (60%) / 9324
serving bug: first sample duplicated	4500+ (30%) / 12434	4029 (50%) / 6217	1771 / 3109
parser bug: 30% answers blank	1319 / 3436	1520 / 1718	545 / 859
serving bug: first two responses misrouted	75 / 106	38 / 53	15 / 27

(c) Escalated LR and delivered disagreement

Condition	Floor 0.05: delay @ paid rate before/after	Delivered disagreement ($p = 0.1$) frozen → fallback
task shift: MATH-500 level ≥ 4	4500+ (40%) @ 0.05→0.11	0.074 → 0.071
serving bug: first sample duplicated	2111 (60%) @ 0.05→0.24	0.088 → 0.065
parser bug: 30% answers blank	518 @ 0.05→0.45	0.143 → 0.043
serving bug: first two responses misrouted	49 @ 0.05→0.50	0.884 → 0.009

Table 45: LLM domain, GSM8K deployment stream (N=64 samples; medians over 10 streams of 3000).

$\bar{p}$	audits to certify α: median over certifying streams (fraction certifying) IPW	score	isotonic	net saving
0.05	118 (40%)	126 (30%)	112 (40%)	0.895
0.1	133 (70%)	133 (70%)	137 (70%)	0.847
0.2	152 (80%)	159 (90%)	189 (90%)	0.752

Table 46: M4Raw: mean audits until certification at $\alpha = 0.05$, and net saving including audits (10 streams of 4000). Parentheses give the fraction certifying when below one; no unsafe candidate certified.

tol. q	λ	risk	$\bar{p}$	audits to certify α (fraction of streams if $< 100\%$)			net saving
				IPW	raw score	isotonic	
0.75	0.2	0.017	0.05	106	110	110	0.294
0.75	0.2	0.017	0.1	113 (90%)	122 (90%)	118 (90%)	0.278
0.75	0.2	0.017	0.2	189	192	189	0.246
0.75	0.2	0.017	0.5	215	215	226	0.152
0.75	0.3	0.042	0.05	– (0%)	121 (10%)	– (0%)	0.308
0.75	0.3	0.042	0.1	77 (10%)	79 (10%)	80 (10%)	0.291
0.75	0.3	0.042	0.2	89 (30%)	206 (30%)	74 (20%)	0.258
0.75	0.3	0.042	0.5	424 (30%)	350 (30%)	425 (30%)	0.159
0.5	0.1	0.017	0.05	68 (50%)	90 (70%)	80 (70%)	0.108
0.5	0.1	0.017	0.1	134 (90%)	119 (90%)	130	0.102
0.5	0.1	0.017	0.2	110	123	111	0.090
0.5	0.1	0.017	0.5	74	73	73	0.054
0.5	0.2	0.042	0.05	64 (20%)	– (0%)	76 (10%)	0.213
0.5	0.2	0.042	0.1	– (0%)	– (0%)	– (0%)	0.201
0.5	0.2	0.042	0.2	113 (10%)	– (0%)	– (0%)	0.178
0.5	0.2	0.042	0.5	77 (10%)	80 (10%)	79 (10%)	0.109
0.5	0.3	0.136	0.05	– (0%)	– (0%)	– (0%)	0.285
0.5	0.3	0.136	0.1	– (0%)	– (0%)	– (0%)	0.269
0.5	0.3	0.136	0.2	– (0%)	– (0%)	– (0%)	0.238
0.5	0.3	0.136	0.5	– (0%)	– (0%)	– (0%)	0.147
0.5	0.4	0.203	0.05	– (0%)	– (0%)	– (0%)	0.308
0.5	0.4	0.203	0.1	– (0%)	– (0%)	– (0%)	0.291
0.5	0.4	0.203	0.2	– (0%)	– (0%)	– (0%)	0.258
0.5	0.4	0.203	0.5	– (0%)	– (0%)	– (0%)	0.159
0.5	0.7	0.212	0.05	– (0%)	– (0%)	– (0%)	0.311
0.5	0.7	0.212	0.1	– (0%)	– (0%)	– (0%)	0.293
0.5	0.7	0.212	0.2	– (0%)	– (0%)	– (0%)	0.260
0.5	0.7	0.212	0.5	– (0%)	– (0%)	– (0%)	0.160

Table 47: Early-output and full-budget task error on held-out GSM8K and Imagenette pools. "Corrects" means full-budget right, early wrong; "spoils" means the reverse. Frozen rules use LLM threshold 0.2 and cascade max-softmax threshold 0.7.

pool	task error, early	task error, full	disagreement	completion corrects	completion harms
LLM (GSM8K-train held-out)					
held-out clean	0.040	0.037	0.003	0.003	0.000
top-p 0.9	0.039	0.038	0.010	0.005	0.004
general model	0.112	0.083	0.043	0.033	0.004
token cap 512	0.040	0.037	0.004	0.003	0.000
token cap 256	0.401	0.380	0.133	0.035	0.014
prompt without boxed	0.039	0.036	0.006	0.004	0.001
cascade (Imagenette)					
clean	0.116	0.111	0.023	0.013	0.009
off-by-one	0.716	0.111	0.615	0.605	0.000
swap	0.288	0.111	0.196	0.186	0.009
pixelate 0.25	0.838	0.916	0.134	0.004	0.082
noise 0.16	0.238	0.202	0.050	0.039	0.004
jpeg 5	0.436	0.433	0.033	0.011	0.009