KNOW WHEN TO TRUST HISTORY: CERTIFYING MODEL UPDATES UNDER EVALUATOR DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Releasing a model update requires certifying that its risk on the current population stays below a threshold. Trusted labels are expensive, whereas a cheap evaluator, such as a language-model judge, scores every example. Reusing evaluator errors from earlier audits is tempting, yet it is unclear when such evidence may replace current labels. The answer depends on the status given to history. If the errors can change invisibly, no label-free test detects the change, and every valid, useful certifier must keep buying labels at a rate we characterize; if a bound on the change is assumed, label-free certification is valid at an explicit error cost. For the middle ground, where history is informative but untrusted, we propose *portfolio vigilance*, a sequential certifier that mixes a betting expert guided by history with one that learns only from current labels; history affects only how it bets, so validity holds for any history. The contribution is not prior-informed betting or expert mixtures themselves, but the distinction between historical evidence that may enter validity and historical evidence that may only guide evidence accumulation. In a canonical model, accurate history shortens decisions but never raises the evidence growth rate, whereas stale history can destroy it. On held-out CIFAR-10N and DICES-990 data, portfolio vigilance needs 0.465 (95% confidence interval (CI) $[0.327, 0.575]$) and 0.740 ($[0.618, 0.877]$) times the labels of a matched prediction-powered monitor with no observed false certification, and fewer labels on all six external blocks. When historical advice is corrupted, it stays within 8.0% of its better component, whereas trusting history alone costs up to 1.66 times as much. In post-confirmatory repeated-judge experiments on DICES-990 and ToxicChat, changing the rubric of a fixed LLM judge moves its scores more than its run-to-run variation. After the change, the portfolio needs 0.790 ($[0.667, 0.909]$) and 0.631 ($[0.520, 0.770]$) times the labels of the matched monitor, and fewer labels than trusting history alone.

1 INTRODUCTION

Suppose a team fine-tunes a deployed model and must decide whether the new version may be released. The question is statistical: does the candidate's risk on the population it serves stay below a threshold, at a stated error level? Trusted labels, such as expert safety ratings, answer it but are slow and expensive. Cheap evidence is abundant: a learned evaluator or a large language model (LLM) judge can score every example, and earlier audits leave a ledger of examples that carry both cheap scores and trusted labels. Prediction-powered evaluation shows how to combine the two (Angelopoulos et al., 2023a). Writing Y for the trusted loss and Q for its cheap prediction, the risk splits into a free part and a paid part,

$$\underbrace{\mathbb{E}_T[Y]}_{\text{target risk}} = \underbrace{\mathbb{E}_T[Q]}_{\text{recomputable without labels}} + \underbrace{\mathbb{E}_T[Y-Q]}_{\text{needs trusted labels}}. \tag{1}$$

Only the mean residual $E = Y - Q$, the evaluator's average error, needs trusted labels. A historical ledger measures exactly this quantity, so reusing it is tempting: the model and the population change, but perhaps the evaluator's errors do not.

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.

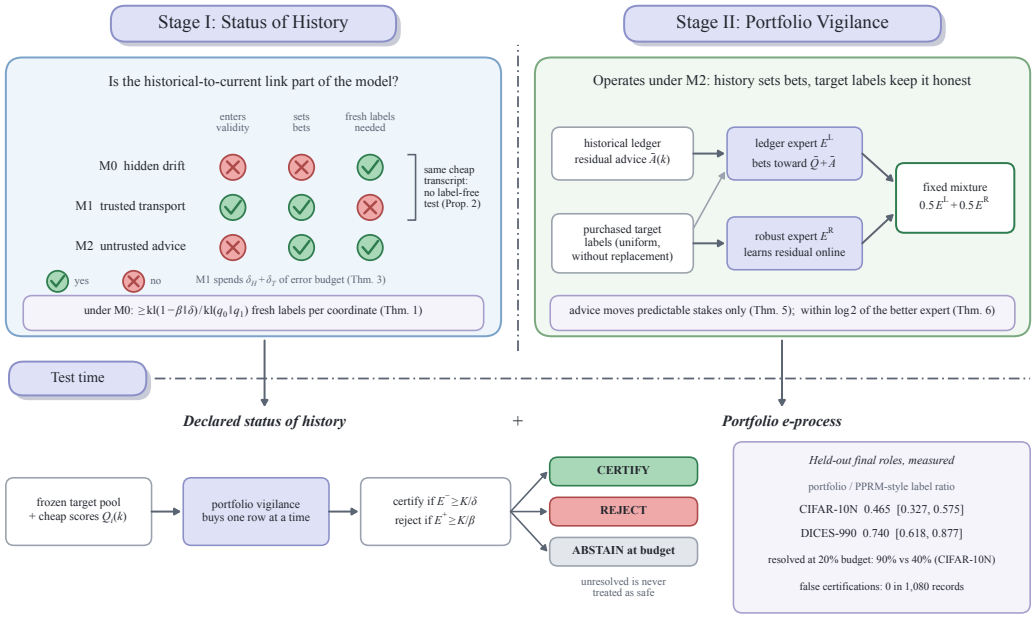


Figure 1: Overview. **Stage I:** hidden drift (M0) and trusted transport (M1) are indistinguishable from cheap data (Proposition 2); M0 forces fresh labels (Theorem 1), while M1 certifies without them at a declared error cost (Theorem 3). **Stage II:** under untrusted advice (M2), a ledger expert and a robust expert bet on each purchased label, and their equal-weight mixture is valid for any advice (Theorem 4) and within $\log 2$ evidence of the better expert (Theorem 5). **Test time:** rows are bought until each candidate is certified, rejected, or unresolved.

Whether this reuse is legitimate is the question of this paper. Existing methods estimate the residual from current labels (Angelopoulos et al., 2023a; Zhang et al., 2026), calibrate the current evaluator against human labels (Feng et al., 2026), or let side information steer anytime-valid evidence without affecting its validity (Waudby-Smith & Ramdas, 2024; Chen & Weng, 2026). None of them says when historical residual evidence may replace current labels and when it may only guide how they are collected. Nor does any certifier capture the benefit of informative but untrusted history while avoiding the cost of stale history, which can be large because changing the evaluator changes the measurement itself (Yang et al., 2026).

We answer both questions by distinguishing three statuses of historical evidence (Figure 1). Under *hidden residual drift* (M0), the evaluator's error can change without any trace in the cheap data, so history need not constrain the present. Under *trusted transport* (M1), a declared bound limits how far the error can move, and history may enter the validity argument directly. Under *untrusted advice* (M2), history may predict the present error but is not assumed correct, so it may only steer how evidence is collected. History can replace current labels only under M1, and because M0 and M1 look identical in cheap data, that status must be assumed, not learned. For M2, we propose *portfolio vigilance*. It samples the target pool at random, runs one betting process guided by history and one that learns the error from purchased labels alone, and certifies from their fixed mixture. Accurate history then saves labels, while stale history costs little and never affects validity.

Our contributions are:

- **Transport or vigilance.** We prove that hidden drift and trusted transport are indistinguishable from cheap data, give a tight lower bound on the fresh labels honest certification needs, and price the error cost of a trusted bridge.
- **A phase diagram for untrusted history.** We characterize when historical advice helps, stops helping, or destroys evidence, and quantify the price of hedging.

- **Portfolio vigilance.** We propose a finite-population certifier that is valid under arbitrary advice and optimal among fixed mixtures in worst-case evidence regret.
- **Evaluation.** In a prespecified evaluation on held-out DICES-990 and CIFAR-10N roles, we compare against matched baselines, with external, corrupted-advice, and sensitivity analyses; a post-confirmatory study with repeated LLM judges on DICES-990 and ToxicChat tests evaluator drift directly.

Prediction-powered inference, e-values, betting confidence sequences, prior-informed betting, and expert mixtures all predate this work; the contribution is the information-model distinction, its consequences, and a certifier that makes it operational.

2 Related work

Prediction-powered and anytime-valid inference. Prediction-powered inference corrects a cheap predictor's mean with labeled residuals (Angelopoulos et al., 2023a;b; Zrnic & Candès, 2024b;a). Its anytime-valid and monitoring variants include prediction-powered e-values (Csillag et al., 2025), prediction-powered risk monitoring (PPRM) (Zhang et al., 2026), and judge auditing for best-arm identification (Ao et al., 2026). Betting confidence sequences (Howard et al., 2021; Waudby-Smith & Ramdas, 2024; 2020; Ville, 1939) underlie risk monitors (Podkopaev & Ramdas, 2022; Timans et al., 2025; Koebler et al., 2025; Xu et al., 2024). Simulator-guided betting also shows that side information can steer the bets of an anytime-valid certificate whose validity holds for any simulator bank (Chen & Weng, 2026). All of these estimate the residual from current labels; none says when a historical residual may replace them.

Changing evaluators. Noisy-but-Valid calibrates a current judge against human labels (Feng et al., 2026). Judge-replacement audits show that changing the judge changes the measurement (Yang et al., 2026; Li, 2026), and disagreement-based auditing exploits observable disagreement between models (Balachandran, 2026). These works study the current evaluator or a visible signal; we ask whether residual evidence from a past evaluator still describes the present one.

Transport and adaptation. Transport results assume a known shift or a declared slack (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021; Doula, 2026; Klivans et al., 2024). Honest procedures, however, cannot adapt to unknown structure for free (Low, 1997; Cai & Low, 2004; Armstrong & Kolesár, 2018; Mazzetto & Upfal, 2023). Our lower bound uses the change-of-measure argument of Kaufmann et al. (2016). We study the case these results leave open: a historical residual whose drift may leave no trace in any cheap observable. Appendix B gives a claim-level comparison.

3 Current-risk certification and information models

Table 3 in the appendix collects the notation for reference. We audit a finite target pool of N rows and K prespecified candidate claims. For row i and candidate k, let $Y_i(k) \in [0,1]$ be a trusted loss and $Q_i(k) \in [0,1]$ a cheap prediction of it, available for the whole pool before any trusted outcome is purchased. The current finite-population risk is

$$R(k) = \frac{1}{N}\sum_{i=1}^{N} Y_i(k), \qquad Y_i(k) = Q_i(k) + E_i(k), \tag{2}$$

where $E_i(k)$ is the current evaluator residual. Candidate k is safe when $R(k) \le \tau_k$, with τ_k fixed in advance. A certifier purchases trusted outcomes one row at a time, and each purchased row updates every candidate defined on it. It must be *valid*, certifying some unsafe candidate with probability at most δ, and should be *useful*, resolving safe candidates with probability at least $1-\beta$. A ledger of paired cheap and trusted outcomes from an earlier role supplies a residual estimate $\bar{A}(k)$ of the historical mean residual $\mu_E^{\mathrm{hist}}(k)$.

M0: unrestricted hidden residual drift. The evaluator's error may change while every cheap observable keeps its distribution. Historical residuals then need not constrain the current residual mean $\mu_E^{\mathrm{tar}}(k)$.

M1: trusted transport. Before target labels are inspected, a historical procedure supplies $U_H(k)$ with $\Pr\{\mu_E^{\mathrm{hist}}(k) \le U_H(k)\} \ge 1-\delta_H$, and a declared transport model supplies a radius ρ_k with $\Pr\{\mu_E^{\mathrm{tar}}(k) \le \mu_E^{\mathrm{hist}}(k) + \rho_k\} \ge 1-\delta_T$. The bridge is part of the model class, so it may enter validity directly. Zero-fresh-label certification is legitimate whenever

$$\overline{Q}_{\mathrm{tar}}(k) + U_H(k) + \rho_k \le \tau_k. \tag{3}$$

As ρ_k grows, this zero-label region shrinks.

M2: untrusted advice. Historical residuals may predict current residuals but are not assumed correct; they may set predictable bets but may not alter the null, the threshold, or the error guarantee.

4 WHY VIGILANCE IS NECESSARY

The difficulty under M0 is that a safe world and an unsafe world can produce identical cheap data, so only trusted labels tell them apart, and a certifier that is valid in both must buy some. We make this precise in a *Bernoulli trusted-loss submodel* of M0 with M monitored coordinates, for example deployment blocks times independently auditable strata. The cheap prediction is identically zero, $Q \equiv 0$, so the residual equals the trusted loss, and each trusted-label request at coordinate i reveals an independent draw $Y \sim \mathrm{Ber}(q)$. The submodel lies in M0 because M0 does not restrict how the residual law changes. The composite safe class has mean at most q_0 and the unsafe class at least q_1, $0 < q_0 < q_1 < 1$. Under the stationary boundary-safe law P_0 every queried loss is $\mathrm{Ber}(q_0)$, and P_i makes coordinate i alone $\mathrm{Ber}(q_1)$. The cheap transcript has the same law under P_0 and every P_i. Let $C_i = 1$ when coordinate i is certified and let N_i be the number of labels queried there. We require validity across all coordinates (honesty) and certification of every coordinate of a safe stream (liveness),

$$\Pr\big(\exists i:\ i \text{ unsafe and } C_i = 1\big) \le \delta, \qquad P_0\big(C_i = 1 \text{ for every } i\big) \ge 1-\beta, \tag{4}$$

with $1-\beta > \delta$ and $\mathrm{kl}(a\|b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$.

Theorem 1 (Price of vigilance)**.** *Every auditor satisfying Eq. 4 obeys, for every coordinate i,*

$$\mathbb{E}_0[N_i] \ \ge\ \frac{\mathrm{kl}(1-\beta\|\delta)}{\mathrm{kl}(q_0\|q_1)}, \qquad \textit{hence} \qquad \mathbb{E}_0\Big[\sum_{i=1}^{M} N_i\Big] \ \ge\ M\,\frac{\mathrm{kl}(1-\beta\|\delta)}{\mathrm{kl}(q_0\|q_1)}. \tag{5}$$

Proof sketch. The all-certified event has probability at least $1-\beta$ under P_0 and at most δ under P_i. Only labels queried at coordinate i distinguish the two laws, so the chain rule for a predictably sampled transcript gives $D_{\mathrm{KL}}(P_0^{\mathrm{tr}}\|P_i^{\mathrm{tr}}) = \mathbb{E}_0[N_i]\,\mathrm{kl}(q_0\|q_1)$, and data processing through the indicator of that event yields the bound.

The cost is charged under P_0, where nothing changes: guarding against a possible change costs labels even when none occurs. A sequential likelihood-ratio certifier attains the bound up to an overshoot constant, and a deterministic cap pays an extra $\log M$ (Appendix E).

Proposition 2 (No label-free validation of transport)**.** *Let Ψ be any transport-validity test measurable with respect to the cheap transcript. Then $P_0(\Psi = 1) = P_i(\Psi = 1)$ for every coordinate i.*

Residual stability therefore cannot be verified from cheap data and must be assumed; Theorem 3 prices that assumption. The failure is not only theoretical: in an exploratory Folktables stress test with hidden injected failures, a certifier that trusts a calibrated bridge without verifying it false-certifies up to 75% of the unsafe blocks, exactly as often as a no-label oracle, while fresh-label baselines catch every one (Appendix L).

Theorem 3 (Trusted-transport certification region)**.** *For each candidate k let $H_k = \{\mu_E^{\mathrm{hist}}(k) \le U_H(k)\}$ and $B_k = \{\mu_E^{\mathrm{tar}}(k) \le \mu_E^{\mathrm{hist}}(k) + \rho_k\}$, with $\Pr(\bigcap_k H_k) \ge 1-\delta_H$ and $\Pr(\bigcap_k B_k) \ge 1-\delta_T$, and define the zero-label slack $s_k = \tau_k - \overline{Q}_{\mathrm{tar}}(k) - U_H(k)$ and the certifiable set $\mathcal{C}(\rho) = \{k:\ s_k \ge \rho_k\}$.*

(i) Certifying every $k \in \mathcal{C}(\rho)$ with zero fresh labels certifies some unsafe candidate with probability at most $\delta_H + \delta_T$. No independence between the historical and bridge events is required.

(ii) If the bridge is part of the model class deterministically ($\delta_T = 0$), the error is at most δ_H, and it is zero if U_H is also exact.

(iii) For a common radius $\rho_k \equiv \rho$, $\mathcal{C}(\rho)$ is nonincreasing in ρ, the zero-label certifiable fraction $|\mathcal{C}(\rho)|/K$ is the empirical survival function of the slacks $\{s_k\}$, and $\mathcal{C}(\rho) = \emptyset$ once $\rho > \max_k s_k$; the limit $\rho \to \infty$ recovers M2, in which no candidate is certifiable without current labels.

The zero-label benefit of M1 also cannot be learned adaptively from the same transcript (Corollary 7, Appendix C).

5 PORTFOLIO VIGILANCE

We now turn to M2. Portfolio vigilance tests each candidate by betting. It holds a wealth that starts at one, multiplies it after each purchased label by a factor whose conditional mean is at most one when the candidate is unsafe, and certifies once the wealth is large. Such a wealth process is an e-process, and Ville's inequality (Ville, 1939) bounds the probability that it ever exceeds $1/\delta$ under the null. Formally, the certifier purchases rows of the target pool in one uniformly random order $\pi_1, \pi_2, \ldots$ without replacement, and before draw t its information $\mathcal{F}_{t-1}$ contains the cheap transcript of the whole pool, the fixed historical advice, and the previously purchased outcomes only. For candidate k let $S_{t-1}(k) = \sum_{s<t} Y_{\pi_s}(k)$ be the cumulative purchased trusted loss. Under the certification null $R(k) \geq \tau_k$, the unpurchased rows have mean at least

$$m_t(k) = \frac{N\tau_k - S_{t-1}(k)}{N - t + 1}, \tag{6}$$

so for any predictable stake $\lambda_t \geq 0$ with $1 + \lambda_t(m_t - y) \geq 0$ on $[0, 1]$, the factor $1 + \lambda_t(k)\{m_t(k) - Y_{\pi_t}(k)\}$ has conditional mean at most one.

Two experts. The *ledger expert* forecasts the remaining-pool mean with the historical correction,

$$\hat{p}^{\mathrm{L}}(k) = \mathrm{clip}\left(\bar{Q}(k) + \bar{A}(k), 0, 1\right),$$

where $\bar{Q}(k)$ is the all-pool cheap mean. The *robust expert* starts with no residual correction and learns only from purchased target labels,

$$\hat{p}^{\mathrm{R}}_t(k) = \mathrm{clip}\left(\bar{Q}(k) + \frac{1}{t-1}\sum_{s<t}\left[Y_{\pi_s}(k) - Q_{\pi_s}(k)\right], 0, 1\right).$$

The ledger expert thus trusts history from the first label, whereas the robust expert must learn the error before it bets confidently. For expert e the certification-direction Kelly stake is $\lambda^{e,-}_t(k) = [m_t(k) - \hat{p}^e_t(k)]_+ / \{m_t(k)(1 - m_t(k))\}$, with the symmetric rejection-direction bet against the null $R(k) \leq \tau_k$. The deployed evidence is the fixed mixture

$$E^-_t(k) = \tfrac{1}{2}E^{\mathrm{L},-}_t(k) + \tfrac{1}{2}E^{\mathrm{R},-}_t(k), \qquad E^+_t(k) = \tfrac{1}{2}E^{\mathrm{L},+}_t(k) + \tfrac{1}{2}E^{\mathrm{R},+}_t(k). \tag{7}$$

Algorithm 1 in Appendix C gives the procedure: a candidate is certified when $E^-_t(k) \geq K/\delta$, rejected when $E^+_t(k) \geq K/\beta$, resolved earlier by exact finite-population bounds, and otherwise left unresolved at the budget, never silently safe.

Theorem 4 (Anytime validity under arbitrary advice)**.** *With losses in $[0, 1]$, uniform sampling without replacement, thresholds and advice fixed without target-role outcomes, and predictable stakes, portfolio vigilance certifies some unsafe candidate with probability at most δ and rejects some safe candidate with probability at most β. The guarantee holds for arbitrary historical advice, candidate-specific stopping, shared purchased outcomes, and abstention at budget exhaustion.*

Proof sketch. Under the null, every predictable betting factor has conditional mean at most one, so each expert's wealth and their mixture are nonnegative supermartingales, and Ville's inequality with a union bound over candidates gives the levels; advice enters only through $\mathcal{F}_{t-1}$. Each assumption is necessary: nonuniform sampling or a stake that looks at the next row can give a betting factor mean above one (Appendix F).

Theorem 5 (Optimal fixed-mixture hedge)**.** *For component e-processes $E_t^{\mathrm{L}}, E_t^{\mathrm{R}}$ and fixed $w \in (0,1)$, $E_t^{\mathrm{mix}} = wE_t^{\mathrm{L}} + (1-w)E_t^{\mathrm{R}}$ is an e-process, and on every path*

$$\log E_t^{\mathrm{mix}} \;\geq\; \max\big\{\log E_t^{\mathrm{L}} + \log w,\; \log E_t^{\mathrm{R}} + \log(1-w)\big\}.$$

The worst-case log-evidence regret to the better component is $c(w) = \max\{-\log w, -\log(1-w)\}$, no smaller uniform constant is possible for that w, and $w = 1/2$ uniquely minimizes $c(w)$ at $\log 2$.

In words, the mixture never trails the better expert by more than a factor of two in evidence; in label units, it reaches any evidence level no later than the first expert reaches twice that level (Corollary 8, Appendix C).

5.1 What history can and cannot buy: a canonical phase diagram

What does it cost *not to know* whether historical residual information is still accurate? A canonical model answers this exactly. Trusted losses of the target block are i.i.d. $\mathrm{Ber}(p)$ with $p < m$, where $m = \tau$ is the certification boundary; this is the infinite-pool limit of Eq. 6. The proxy has constant mean $\bar{Q}$, so $p = \bar{Q} + \mu_E^{\mathrm{tar}}$, and the ledger supplies μ_E^{hist}. The *advice mismatch*

$$\eta = \mu_E^{\mathrm{tar}} - \mu_E^{\mathrm{hist}} \tag{8}$$

turns the ledger forecast into $r = \bar{Q} + \mu_E^{\mathrm{hist}} = p - \eta$. Positive η is optimistic stale advice (history understates current risk) and negative η is pessimistic stale advice. The robust expert forecasts with the running mean of purchased losses, truncated below at a constant $\varepsilon \in (0, p]$ for the analysis.

Proposition 6 (Transport-vigilance phase diagram)**.** *In the canonical model, almost surely:*

(i) the ledger expert's log evidence grows at rate $\lim_t t^{-1}\log E_t^{\mathrm{L}} = G_L(\eta)$, where

$$G_L(\eta) = \begin{cases} 0, & \eta \leq \eta_- := p - m \quad \textit{(forecast at or above the boundary: no bet)}, \\ \mathrm{kl}(p\|m) - \mathrm{kl}(p\|p-\eta), & \eta_- < \eta < p, \end{cases}$$

and $E_t^{\mathrm{L}} \to 0$ if $\eta \geq p$ (a forecast clipped at zero is ruined by the first unit loss);

(ii) G_L is uniquely maximized at $\eta = 0$ with value $\mathrm{kl}(p\|m)$, is positive exactly on (η_-, η_+), where $\eta_+ = p - r_\star \in (0,p)$ and $r_\star \in (0,p)$ is the unique solution of $\mathrm{kl}(p\|r_\star) = \mathrm{kl}(p\|m)$, and is negative for $\eta > \eta_+$;

(iii) the robust expert's rate is $\lim_t t^{-1}\log E_t^{\mathrm{R}} = \mathrm{kl}(p\|m)$ for every η;

(iv) the portfolio's rate is $\lim_t t^{-1}\log E_t^{\mathrm{mix}} = \max\{G_L(\eta), \mathrm{kl}(p\|m)\} = \mathrm{kl}(p\|m)$.

History therefore shortens a decision but never raises the evidence rate; pessimistic stale advice costs the ledger only power, optimistic stale advice beyond η_+ makes its evidence decay, and the portfolio keeps the vigilance rate for every η (Figure 2). Item (i) follows from an exact one-step identity for the ledger's log growth (Theorem 10 in Appendix F). The price of not knowing η is explicit: the portfolio needs at most $\log 2/G_L(\eta)$ more expected labels than the ledger's bound, and at $\eta = 0$ its bound matches the lower bound of Theorem 1 to first order in $\log(1/\delta)$ (Corollary 9, Appendix C).

6 Experimental design

Protocol. To keep analysis choices independent of the results, we fix everything before analyzing the final roles: five random sampling orders (seeds), block-level $\delta = 0.05$ and $\beta = 0.10$, budgets from 1% to 100% of each block, thresholds of ledger risk plus 0.05, and the comparators. All methods share pools, orders, and budgets, and development uses separate datasets.

Data. DICES-990 contains 72,103 safety ratings of 990 conversations (Aroyo et al., 2023); we audit four outcome-level candidates (harmful content, unfair bias, misinformation, and overall safety) on 21,702 held-out rating rows in 12 locale-by-hash blocks, with a term-frequency inverse-document-frequency (TF-IDF) logistic proxy. CIFAR-10N provides three human annotations per CIFAR-10 training image (Wei et al., 2022); the loss is disagreement with the canonical class, and 14,767 held-out images form six blocks with one candidate each. Both are split by hash into proxy-training, ledger, calibration, and final roles (Appendix G), and each final-role block is one target pool.

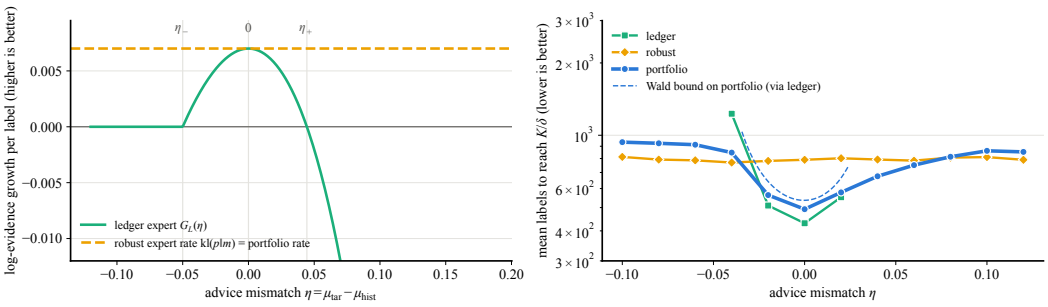


Figure 2: Canonical phase diagram ($p = 0.20$, $m = 0.25$, $K = 1$, $\delta = 0.05$). **Left (exact):** ledger growth $G_L(\eta)$ against the flat robust and portfolio rate $\mathrm{kl}(p\|m)$. **Right (simulation, 2,000 paths per point):** mean labels to reach K/δ; ledger points are omitted where over 1% of paths fail within 6,000 labels, and the dashed curve is the bound of Corollary 9.

Comparators and statistics. We compare the portfolio with its two components (ledger-only and robust-only), fresh without-replacement (WoR) betting, a selective residual confidence sequence (CS) (Ao et al., 2026), passive fixed-budget sampling, and a PPRM-style conjugate-mixture empirical-Bernstein (CM-EB) monitor. The last keeps the prediction-plus-residual estimator of Zhang et al. (2026) but is adapted to our fixed-pool estimand; we do not claim it reproduces the original procedure exactly (Appendix G). The primary metric is the paired geometric-mean portfolio/PPRM label ratio at full budget, with 95% hierarchical bootstrap intervals over blocks and then seeds, and the area under the correct-resolution-versus-budget curve (AUC). In tables ↑ means higher is better and ↓ lower is better.

Post-confirmatory repeated-judge study. Designed after the primary protocol was fixed, this study replaces the cheap proxy with a legacy rubric (Judge A) and an updated contextual-risk rubric (Judge B) of one fixed LLM, so it isolates evaluator-configuration drift rather than a model update. Each rubric judges every item three times, on all 990 DICES conversations and all 5,654 human-annotated ToxicChat-0124 examples (Lin et al., 2023); Judge A forms the ledger and Judge B the current proxy in $A \to B$. The analysis uses the five prespecified primary orders rather than the 50 planned in the written post-confirmatory protocol, averages them within block before 10,000 block-bootstrap replicates, and never treats them as independent units (Appendix K).

7 Results

7.1 Held-out natural data

Table 1 gives the full-budget outcomes. The paired portfolio/PPRM label ratio is 0.465 (95% CI $[0.327, 0.575]$) on CIFAR-10N and 0.740 ($[0.618, 0.877]$) on DICES-990, a reduction of 53.5% and 26.0%. The ledger-only expert is slightly cheaper because these data favor historical advice; the portfolio pays a small hedge for not knowing this, and beats every other deployable baseline. When history is stale the order reverses: after an LLM judge's rubric changes, the portfolio needs fewer labels than ledger-only on DICES-990 and ToxicChat (Section 7.3), and under biased advice ledger-only needs up to 1.66 times the labels of the cheaper component (Section 7.2). The portfolio makes no false certification; its one DICES false rejection, shared by the robust-only expert, is a safe candidate rejected early in one block. Zero observed false certifications are an empirical observation, not evidence of zero error probability; validity follows from the stated assumptions and procedure (Theorem 4).

The advantage holds across the budget frontier (Figure 3 in Appendix H): resolution AUC is 0.880 against 0.743 on CIFAR-10N and 0.569 against 0.402 on DICES-990, and at 20% of the budget the portfolio resolves 90% of CIFAR-10N decisions against 40% and 34.58% of DICES-990 decisions against 13.75%. The DICES gain varies across blocks, from 0.422 to 0.999 (Appendix H).

Table 1: Full-budget results on the held-out final roles. Ledger-only and robust-only are the portfolio's components. FC/FR: false certifications/rejections over 30 (CIFAR-10N) and 240 (DICES-990) decisions; desc.: descriptive. Labels are mean ± standard deviation (SD) over block-seed pairs; comparisons use the paired ratios in the text. Bold marks the lowest labels and highest AUC within each dataset. Selective residual CS buys labels on the passive schedule but resolves nothing before full budget, so its AUC is only the last budget interval, 0.100.

Dataset	Method	Labels ↓	Resolution ↑	Release (desc.)	FC ↓	FR ↓	AUC ↑
CIFAR-10N	**Portfolio vigilance**	279.9 ± 171.0	100.00%	100.00%	0	0	0.880
	Ledger-only (component)	$\mathbf{248.9} \pm 166.5$	100.00%	100.00%	0	0	**0.894**
	Robust-only (component)	967.8 ± 675.5	100.00%	100.00%	0	0	0.602
	PPRM-style CM-EB	574.6 ± 325.7	100.00%	100.00%	0	0	0.743
	Fresh WoR betting	$1{,}492.6 \pm 438.0$	100.00%	100.00%	0	0	0.400
	Passive fixed budget	$2{,}461.2 \pm 39.0$	100.00%	100.00%	0	0	0.840
	Selective residual CS	$2{,}461.2 \pm 39.0$	100.00%	100.00%	0	0	0.100
DICES-990	**Portfolio vigilance**	$1{,}182.2 \pm 678.2$	99.58%	72.50%	0	1	**0.569**
	Ledger-only (component)	$\mathbf{1{,}181.4} \pm 689.3$	100.00%	72.92%	0	0	0.534
	Robust-only (component)	$1{,}368.8 \pm 527.8$	99.58%	72.50%	0	1	0.501
	PPRM-style CM-EB	$1{,}398.3 \pm 542.0$	100.00%	72.92%	0	0	0.402
	Fresh WoR betting	$1{,}586.1 \pm 295.1$	99.17%	72.92%	1	1	0.316
	Passive fixed budget	$1{,}808.5 \pm 176.6$	100.00%	72.92%	0	0	0.521
	Selective residual CS	$1{,}808.5 \pm 176.6$	100.00%	72.92%	0	0	0.100

7.2 External blocks and corrupted advice

On six external blocks from CIFAR-10 with controlled noise, SVHN (Krizhevsky, 2009; Netzer et al., 2011), and UCI Naval Propulsion (Coraddu et al., 2016), the unchanged portfolio uses fewer labels than the PPRM-style monitor on all six (ratios 0.296 to 0.916), with no false certification in any study (Appendix H).

Because the natural data favor accurate history, the hedge is best tested where history is wrong. We therefore bias the historical advice on synthetic finite pools, holding everything else fixed. With exact advice the ledger-only, portfolio, and robust-only experts use 353.4, 381.7, and 594.5 labels; under a -0.15 bias they use 986.7, 624.5, and 594.5. Across seven biases the portfolio stays within 8.0% of the cheaper component, while the ledger-only expert reaches 1.66 times it (Appendix I), which is the finite-pool counterpart of Figure 2.

7.3 Sensitivity and evaluator-shift experiments

Sensitivity. The advantage does not hinge on the prespecified threshold: margins from 0.02 to 0.10 keep the ratio between 0.403 and 0.465 on CIFAR-10N and between 0.649 and 0.740 on DICES-990 (Appendix J).

LLM evaluator shift. The post-confirmatory repeated-judge study (Section 6) tests the motivating mechanism directly. First, the rubric change is real rather than judge noise: the mean absolute score change between rubrics exceeds that within a rubric by 0.0476 (95% CI $[0.0331, 0.0633]$) on DICES-990 final items and by 0.0379 ($[0.0301, 0.0460]$) on ToxicChat. Second, after the change (Table 2) the portfolio needs 0.790 ($[0.667, 0.909]$) and 0.631 ($[0.520, 0.770]$) times the labels of the PPRM-style monitor, with no false certification. Third, unlike on the stable primary data, stale history now makes the ledger expert the costlier one: the portfolio needs 140.7 fewer labels than ledger-only on DICES-990 ($[-256.2, -46.2]$ for the difference) and 12.7 fewer on ToxicChat ($[-20.8, -5.9]$), and it is cheaper than ledger-only in all nine pairings of an A run with a B run on both datasets. This is the hedge that M2 predicts. The portfolio also beats robust-only on DICES-990 (-109.1, $[-199.3, -32.6]$); on ToxicChat the two tie (137.6 against 137.5 labels; difference $+0.05$, $[-4.65, 4.05]$), so we claim no advantage there. Charging the full block for every error or unresolved decision leaves these conclusions unchanged. For comparison, the calibration-based endpoint test Noisy-but-Valid (Feng et al., 2026) makes no false certification but also certifies none

Table 2: Repeated-judge study under the shifted $A \to B$ condition at full budget. Labels are means over blocks and the five random orders. P/P: portfolio/PPRM-style label ratio with its block-bootstrap 95% interval. Shift excess: between-rubric minus within-rubric mean absolute judge-score change on final items, with its item-bootstrap 95% interval. The portfolio makes no false certification on either dataset.

Dataset	Portfolio ↓	Ledger-only ↓	Robust-only ↓	PPRM ↓	P/P ratio ↓	Shift excess
DICES-990	**1,305.2**	1,445.9	1,414.2	1,544.3	0.790 [0.667, 0.909]	0.0476 [0.0331, 0.0633]
ToxicChat	137.6	150.2	**137.5**	208.5	0.631 [0.520, 0.770]	0.0379 [0.0301, 0.0460]

of the eight safe endpoint cases (Appendix K). On DICES-990 the ratio to PPRM is no smaller than in the primary study (0.740): the shift shows that the hedge protects the portfolio, not that it widens its lead over PPRM. An initial single-run study on DICES-990 gives the same pattern (ratio 0.779, $[0.629, 0.924]$; Appendix K).

The price of trusted transport. The repeated-judge data also illustrate M1 (Appendix K.2): at $\rho = 0$ a zero-label certificate would cover 81% of DICES-990 cells under rubric A although the bridge holds for only 35%, and the smallest radius at which the bridge holds everywhere leaves at most 2% certifiable, so a bridge wide enough to be true removes the zero-label advantage.

8 Limitations

The lower bound concerns i.i.d. Bernoulli draws rather than finite pools sampled without replacement, and in a scaling study the portfolio uses four to five times the bound (Appendix I); Proposition 6 and Corollary 9 hold only in the canonical model. Validity rests on the declared design of bounded losses, prespecified thresholds, and uniform sampling: if audited rows are chosen by convenience, a false certificate can release a harmful update whose cost falls on its users. Error control is block-level. On the empirical side, the PPRM-style monitor is our adaptation of the original method. The LLM studies change the rubric of one fixed model, not the model itself, and were designed after the primary protocol was fixed; the repeated-judge study uses the five primary random orders rather than the 50 its written protocol planned, and one of its six comparisons (portfolio against robust-only on ToxicChat) is a tie (Appendix K). The primary label counts do not penalize incorrect decisions, although a penalized cost leaves the repeated-judge conclusions unchanged.

9 Conclusion

Historical evaluator evidence has two statistically different roles. If a trusted transport condition connects an old audit to the present, the audit may constrain current risk and replace fresh labels, at an error cost that should be declared; if not, no label-free test can tell whether the connection still holds, and honest certification must keep buying current evidence at a rate we characterize. In the second regime history can still shorten a decision, but it cannot raise the rate at which evidence accumulates, and stale history can destroy it. Portfolio vigilance pays a small hedge when history is reliable and avoids the larger loss when it is stale: it needs markedly fewer trusted labels than a matched prediction-powered monitor and, when an LLM judge's rubric changes, fewer than trusting history alone, consistent with the M2 mechanism. A release process should therefore state, for every certificate, whether historical evidence entered its validity or only its efficiency. Open questions are how to certify transport itself with few labels when the bridge is only partially testable, and the price of vigilance for finite pools sampled without replacement.

AI use statement

The authors use generative AI tools to assist with literature search and summarization, methodological critique, software design, code drafting and debugging, interpretation of statistical results, figure and table preparation, and manuscript drafting and editing. The authors check every AI-assisted

mathematical and empirical claim against the code, the stored results, the primary literature, and reproducible analyses. The authors take responsibility for the final text, proofs, code, and reported results.

Ethics statement

All experiments use public datasets under their published terms. DICES contains sensitive conversational safety content and socially situated human judgments (Aroyo et al., 2023), CIFAR-10N contains crowd annotations (Wei et al., 2022), and ToxicChat contains real user prompts, some of them toxic, with human toxicity labels (Lin et al., 2023). We report only declared statistical losses and do not treat demographic disagreement as a single objective notion of safety; the locale heterogeneity in Section 7.1 is descriptive. No model or threshold in this paper should be read as a recommendation about people. Results are benchmark measurements, not safety, clinical, or fairness guarantees. The study collects no new human-subject data.

Reproducibility statement

The appendix contains complete proofs, the full per-method and per-block result tables, and the design of every study, including the data roles, random orders, budgets, thresholds, and comparators. The primary protocol is fixed before any analysis of the final roles, and every analysis designed afterwards is marked as additional or post-confirmatory. The supplementary material contains the method and baseline code, the deterministic role construction, the frozen LLM-judge outputs, the per-block results, and scripts that regenerate every table and figure. All experiments run on CPUs.

References

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with LLM judges and limited human. *arXiv preprint arXiv:2601.21471*, 2026.

Timothy B. Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES dataset: Diversity in conversational AI evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

Vishnu Bindu Balachandran. Pay only for disagreement: Certified no-regression verdicts for model updates with matching label-complexity bounds. *arXiv preprint arXiv:2609.17560*, 2026.

T. Tony Cai and Mark G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2004.

Yujia Chen and Bowen Weng. Betting for sim-to-real performance certificates. *arXiv preprint arXiv:2608.21572*, 2026.

Andrea Coraddu, Luca Oneto, Alessandro Ghio, Stefano Savio, Davide Anguita, and Massimo Figari. Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 230(1):136–153, 2016.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11493–11514. PMLR, 2025.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Achref Doula. Conformal calibration transfer. In *Forty-third International Conference on Machine Learning*, 2026.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations*, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Adam R. Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *Conference on Learning Theory*, 2024.

Alexander Koebler, Thomas Decker, Ingo Thon, Volker Tresp, and Florian Buettner. Incremental uncertainty-aware performance monitoring with active labeling intervention. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2188–2196, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yitao Li. Who drifted: the system or the judge? Anytime-valid attribution in LLM evaluation pipelines. *arXiv preprint arXiv:2606.15474*, 2026.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4694–4702, 2023. doi: 10.18653/v1/2023.findings-emnlp.311.

Mark G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

Alessio Mazzetto and Eli Upfal. An adaptive algorithm for learning with unknown distribution drift. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, 2021.

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

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

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A. Naesseth. On continuous monitoring of risk violations under unknown shift. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4204–4226, 2025.

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

Abraham Wald. *Sequential Analysis*. John Wiley and Sons, New York, 1947.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in neural information processing systems*, 33:20204–20214, 2020.

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.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1920.

Zongyou Yang, Yinghan Hou, and Xiaokun Yang. When the judge changes, so does the measurement: Auditing LLM-as-judge reliability. *arXiv preprint arXiv:2607.08535*, 2026.

Guangyi Zhang, Yunlong Cai, Guanding Yu, and Osvaldo Simeone. Prediction-powered risk monitoring of deployed models for detecting harmful distribution shifts. In *Forty-third International Conference on Machine Learning*, 2026.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *International Conference on Machine Learning*, 2024a.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15), 2024b.

The appendix follows the order of the paper. Appendix A summarizes the notation and Appendix B extends the related work. Appendix C gives the pseudocode and the corollaries stated in brief in the main text, Appendix D proves every result of Sections 4 and 5, Appendix E gives the matching sequential certifier and the cost of deterministic label caps, and Appendix F discusses the scope of the theory. Appendix G describes the data roles and the comparator. Appendices H to K report additional results in the order of Section 7: per-block and external results, corrupted advice and scaling, natural-data sensitivity, and the single-run and repeated-judge LLM evaluator-configuration studies. Appendix L reports the Folktables hidden-failure study of unverified transport cited in Section 4.

A Notation

Table 3 summarizes the notation used throughout the paper, grouped in the order in which the symbols appear.

B Extended related work

Table 4 compares the paper with the closest prior results at the level of their guarantees. Prediction-powered e-values give prediction-assisted counterparts of e-value procedures (Csillag et al., 2025), and simulator-guided betting builds anytime-valid certificates whose validity holds for any simulator bank while trust shifts across a portfolio of simulators (Chen & Weng, 2026), so using untrusted side information, or a portfolio over its sources, to choose bets is not new. Noisy-but-Valid estimates a current judge's true- and false-positive rates from human calibration labels and propagates that uncertainty into a valid test (Feng et al., 2026); its problem is judge imperfection, whereas ours is whether residual evidence from a historical judge or population still describes the current one. Disagreement-based auditing proves matching label complexity from observable model disagreement (Balachandran, 2026), whereas the hidden residual changes we study may leave no trace in any cheap observable.

C Algorithm and deferred statements

This appendix collects the pseudocode and three corollaries stated in brief in the main text; their proofs are in Appendix D.

Table 3: Notation.

Symbol	Meaning
M0, M1, M2	Hidden residual drift; trusted transport; untrusted advice
N, K	Rows in the finite target pool; number of prespecified candidate claims
$Y_i(k)$, $Q_i(k)$	Trusted loss and cheap prediction of it for row i and candidate k, both in $[0,1]$
$E_i(k)$	Evaluator residual $Y_i(k) - Q_i(k)$
$R(k)$, τ_k	Current finite-population risk of candidate k; its release threshold
δ, β	False-certification and false-rejection levels
$\mu_E^{\mathrm{hist}}(k)$, $\mu_E^{\mathrm{tar}}(k)$	Historical and current (target) mean residual
$\bar{A}(k)$	Historical residual advice supplied by the ledger
$\bar{Q}(k)$, $\overline{Q}_{\mathrm{tar}}(k)$	All-pool mean of the cheap prediction
$U_H(k)$, δ_H	Historical upper bound on $\mu_E^{\mathrm{hist}}(k)$ and its error probability
ρ_k, δ_T	Trusted-transport (bridge) radius and its error probability
s_k, $\mathcal{C}(\rho)$	Zero-label slack $\tau_k - \overline{Q}_{\mathrm{tar}}(k) - U_H(k)$; candidates certifiable with zero labels
Ψ	Any test measurable with respect to the cheap transcript
M	Monitored coordinates in the lower-bound model (e.g. blocks times strata)
q_0, q_1	Safe and unsafe Bernoulli trusted-loss means in the lower-bound submodel
P_0, P_i	Stationary safe law; law with coordinate i made unsafe
C_i, N_i	Certification indicator and number of labels queried at coordinate i
$\mathrm{kl}(a\|b)$	Binary relative entropy
π_t, $\mathcal{F}_{t-1}$	Row purchased at step t; information available before draw t
$S_{t-1}(k)$, $m_t(k)$	Cumulative purchased trusted loss; null-boundary mean of the unpurchased rows
$\hat{p}^{\mathrm{L}}(k)$, $\hat{p}_t^{\mathrm{R}}(k)$	Ledger-expert and robust-expert forecasts
$\lambda_t^{e,\pm}(k)$	Kelly stake of expert e in the certification $(-)$ or rejection $(+)$ direction
$E_t^{\mathrm{L},\pm}$, $E_t^{\mathrm{R},\pm}$	Wealth (e-process) of the ledger and robust experts
$E_t^{\pm}$, E_t^{mix}	Portfolio evidence, the fixed mixture of the two experts
w, $c(w)$	Ledger mixture weight; worst-case log-evidence regret to the better expert
A, $T_e(A)$	Evidence level; first purchased-label time at which expert e reaches A
p, m	Current trusted-loss mean and certification boundary in the canonical model
r, $B_r(Y)$	Ledger forecast $p-\eta$; one-step Kelly factor with forecast r
η	Advice mismatch $\mu_E^{\mathrm{tar}} - \mu_E^{\mathrm{hist}}$
η_-, η_+, $r_\star$	Phase boundaries $p-m$ and $p-r_\star$; solution of $\mathrm{kl}(p\|r_\star) = \mathrm{kl}(p\|m)$ below p
$G_L(\eta)$	Ledger expert's expected log-evidence growth per label
ε, c_η	Truncation of the robust forecast; largest ledger log-wealth increment

Corollary 7 (No free adaptation to an unverified bridge)**.** *If a safe world and an unsafe M0 world induce the same pre-label transcript, any procedure that is uniformly δ-valid over M0 and decides from that transcript to issue a zero-label transport certificate does so in the safe world with probability at most δ.*

Corollary 8 (Pathwise stopping envelope)**.** *Let $T_{\mathrm{mix}}(A)$, $T_{\mathrm{L}}(A)$, and $T_{\mathrm{R}}(A)$ be the first purchased-label times at which the mixture, ledger, and robust evidence reach level A. On every path $T_{\mathrm{mix}}(A) \le \min\{T_{\mathrm{L}}(A/w),\, T_{\mathrm{R}}(A/(1-w))\}$; with $w = 1/2$ the portfolio reaches A no later than the first component reaches $2A$.*

Corollary 9 (Price of hedging)**.** *In the canonical model with $\eta \in (\eta_-, \eta_+)$, let T_{port} be the first label at which the portfolio's certification evidence reaches K/δ. Then*

$$\mathbb{E}[T_{\mathrm{port}}] \;\le\; \frac{\log(2K/\delta) + c_\eta}{G_L(\eta)}, \qquad c_\eta = \log\frac{1-p+\eta}{1-m},$$

while the ledger expert alone satisfies the same bound with $\log(K/\delta)$ in place of $\log(2K/\delta)$. Not knowing whether history is accurate therefore costs at most $\log 2/G_L(\eta)$ additional expected labels relative to the ledger's bound, and at $\eta = 0$ the portfolio's bound is $\{\log(2K/\delta) + c_0\}/\,\mathrm{kl}(p\|m)$, which matches the lower bound $\mathrm{kl}(1-\beta\|\delta)/\,\mathrm{kl}(q_0\|q_1)$ of Theorem 1, taken with $q_0 = p$ and $q_1 \downarrow m$, to first order in $\log(1/\delta)$ up to the factor $1-\beta$ and additive constants.

Table 4: Positioning against the closest prior results; guarantees are not numerically comparable.

Work	Guarantee	What this paper adds
Zhang et al. (2026)	Anytime monitoring of prediction-powered running-average risk under shift.	Current-risk release decisions and the status of a historical residual ledger.
Feng et al. (2026)	Valid endpoint tests with an imperfect current judge.	Whether historical residuals may enter validity after the judge changes.
Chen & Weng (2026)	Anytime-valid confidence sequence for a mean from i.i.d. real outcomes, valid for any simulator bank; exponential-weights trust combines simulator forecasts into one bet, with wealth-regret bounds against the best fixed simulator.	A fixed mixture of the wealths of a history-guided and a label-only expert, within $\log 2$ evidence of the better one (minimax constant), a stale-advice growth identity, and finite-pool release decisions.
Xu et al. (2024)	Side information guides betting without entering validity.	A two-expert fixed mixture with an optimal $\log 2$ hedge and a stale-advice growth identity.
Csillag et al. (2025)	Prediction-powered counterparts of e-value procedures.	Historical, not current, prediction residuals, and when they may enter validity.
Balachandran (2026)	Matching label complexity for update auditing from observable disagreement.	A rate driven by residual transport and its untestable part, not by disagreement.
Mazzetto & Upfal (2023)	Latest-distribution estimation under unknown drift.	Selective label counting, a release certificate, and a cost charged on a stationary path.
Low (1997); Cai & Low (2004)	Non-adaptivity of honest confidence intervals.	A sequential selective-label vigilance rate and a transport condition that escapes it.
Doula (2026); Klivans et al. (2024)	Transfer or learning under shift with a surrogate or testable condition.	Impossibility of label-free validation of the residual-transport contract.

Algorithm 1 Two-sided portfolio vigilance for one finite target block

Require: pool of N rows with cheap scores $Q_i(k)$; thresholds τ_k; fixed advice $\bar{A}(k)$; levels δ, β; label budget B
Ensure: a decision in {certify, reject, abstain} for each of the K candidates
Draw a uniform random order π of the pool; set all expert wealths to 1
for $t = 1, \ldots, B$ **do**
  if every candidate is resolved **then**
    stop
  end if
  for each unresolved candidate k **do**
    compute $m_t(k)$ by Eq. 6, forecasts $\hat{p}^{\mathrm{L}}(k)$ and $\hat{p}^{\mathrm{R}}_t(k)$ from $\mathcal{F}_{t-1}$, and the four directional Kelly stakes {predictable: never uses row π_t}
  end for
  purchase $Y_{\pi_t}(k)$ for every candidate defined on row π_t {one label updates all candidates}
  update the four expert wealths of each unresolved candidate and form $E^{\pm}_t(k)$ by Eq. 7
  certify k if $E^{-}_t(k) \geq K/\delta$ or $(S_t + N - t)/N \leq \tau_k$; reject k if $E^{+}_t(k) \geq K/\beta$ or $S_t/N > \tau_k$
end for
report unresolved candidates as abstentions

D PROOFS

This section proves the results of the main text in the order in which they appear. Throughout, P^{tr} denotes the law of the observed transcript, which contains every cheap observation, every purchased label, and every internal decision of the auditor.

D.1 Proofs for Section 4

Proof of Theorem 1. Let $A = \{C_i = 1 \text{ for every } i\}$. Liveness gives $P_0(A) \geq 1 - \beta$. Under P_i coordinate i is unsafe, so certifying it violates honesty and $P_i(A) \leq \delta$. Only labels queried at coordinate i have different laws under P_0 and P_i. If $\mathbb{E}_0[N_i] = \infty$ the claim is immediate. Otherwise apply the chain rule for relative entropy to the transcript stopped after the first r queries at coordinate i. Because each query decision is predictable given the past transcript, the per-step contribution is $\mathrm{kl}(q_0 \| q_1)$ when a label at coordinate i is drawn and zero otherwise, so $D_{\mathrm{KL}}(P_0^{\mathrm{tr}} \| P_i^{\mathrm{tr}}) = \mathbb{E}_0[N_i \wedge r] \, \mathrm{kl}(q_0 \| q_1)$ for the stopped transcript. Letting $r \to \infty$ and applying monotone convergence gives $D_{\mathrm{KL}}(P_0^{\mathrm{tr}} \| P_i^{\mathrm{tr}}) = \mathbb{E}_0[N_i] \, \mathrm{kl}(q_0 \| q_1)$. Data processing through the indicator of A gives

$$\mathbb{E}_0[N_i] \, \mathrm{kl}(q_0 \| q_1) \;\geq\; \mathrm{kl}\big(P_0(A) \,\|\, P_i(A)\big) \;\geq\; \mathrm{kl}(1 - \beta \| \delta), \tag{9}$$

using that $\mathrm{kl}(a \| b)$ is increasing in a above b and decreasing in b below a, with $1 - \beta > \delta$. Summing over i gives the cumulative bound. □

A by-product is worth recording. On the event $N_i = 0$ the transcript has the same law under both measures, so $P_0(C_i = 1, N_i = 0) = P_i(C_i = 1, N_i = 0) \leq \delta$, and subtracting from $P_0(C_i = 1) \geq 1 - \beta$ gives $P_0(N_i \geq 1) \geq 1 - \beta - \delta$. Vigilance is not only expensive in expectation; with high probability it happens at every coordinate.

Proof of Proposition 2. The cheap transcript has the same law under P_0 and P_i by construction, and Ψ is measurable with respect to it, so Ψ has the same law under both. □

Proof of Theorem 3. (i) Let $H = \bigcap_k H_k$ and $B = \bigcap_k B_k$. On $H \cap B$, every k satisfies $\mu_E^{\mathrm{tar}}(k) \leq \mu_E^{\mathrm{hist}}(k) + \rho_k \leq U_H(k) + \rho_k$. For $k \in \mathcal{C}(\rho)$, $s_k \geq \rho_k$ means $\overline{Q}_{\mathrm{tar}}(k) + U_H(k) + \rho_k \leq \tau_k$, so $R(k) = \overline{Q}_{\mathrm{tar}}(k) + \mu_E^{\mathrm{tar}}(k) \leq \tau_k$ and k is safe. A false zero-label certificate therefore requires $H^c \cup B^c$, and $\Pr(H^c \cup B^c) \leq \delta_H + \delta_T$ by a union bound, which needs no independence. The proxy mean $\overline{Q}_{\mathrm{tar}}(k)$ is computed exactly on the whole pool and contributes no error. (ii) With $\delta_T = 0$ the bound reads δ_H; if U_H also holds surely, $H^c \cup B^c$ is null. (iii) $k \in \mathcal{C}(\rho)$ iff $s_k \geq \rho$, so $\rho \leq \rho'$ implies $\mathcal{C}(\rho') \subseteq \mathcal{C}(\rho)$, and $|\mathcal{C}(\rho)|/K = K^{-1} \sum_k \mathbf{1}\{s_k \geq \rho\}$ is the empirical survival function of the slacks, which vanishes for $\rho > \max_k s_k$. □

Proof of Corollary 7. Let $\mathcal{G}_0$ be the pre-label transcript and $A \in \mathcal{G}_0$ the event that a zero-label certificate is issued. The two worlds restrict to the same law on $\mathcal{G}_0$, so $P_{\mathrm{safe}}(A) = P_{\mathrm{unsafe}}(A) \leq \delta$. Any rule that tries to infer from the same transcript whether to trust a bridge is still $\mathcal{G}_0$-measurable and is covered by the same equality. □

D.2 Proofs for Section 5

Proof of Theorem 4. Fix a candidate and suppress k. Under $R \geq \tau$, the rows not yet drawn sum to $NR - S_{t-1} \geq N\tau - S_{t-1}$, so uniform sampling from them gives $\mathbb{E}[Y_{\pi_t} \mid \mathcal{F}_{t-1}] \geq m_t$ with m_t as in Eq. 6. For predictable $\lambda_t \geq 0$ with $1 + \lambda_t(m_t - y) \geq 0$ for all $y \in [0, 1]$, $\mathbb{E}[1 + \lambda_t(m_t - Y_{\pi_t}) \mid \mathcal{F}_{t-1}] \leq 1$, so each expert's wealth is a nonnegative supermartingale with unit start, and so is any fixed convex mixture of the two. Ville's inequality bounds the probability that the certification mixture of a given unsafe candidate ever reaches K/δ by δ/K, and a union bound over the at most K candidates gives the family-wise false-certification bound δ. The rejection direction reverses the centered factor under $R \leq \tau$ and gives β in the same way. The stakes use only $\mathcal{F}_{t-1}$, which contains the historical advice as a fixed quantity; no property of the advice is used, which is why the guarantee holds for arbitrary advice. When $m_t \notin (0, 1)$ the implementation skips betting and relies on the deterministic closure bounds $L_t = S_t/N$ and $U_t = (S_t + N - t)/N$, which are exact and so preserve validity. One purchased row updates every candidate, stopping one candidate does not change the random-order design for the others, and unresolved candidates at budget exhaustion are abstentions. □

A stake that conditions on the proxy of the next selected row is not $\mathcal{F}_{t-1}$-measurable and breaks the supermartingale property; the certifier therefore uses only all-pool cheap summaries and previously purchased outcomes.

Proof of Theorem 5. A fixed convex combination of nonnegative supermartingales with unit start is again one, so E^{mix} is an e-process. Pathwise, $E_t^{\mathrm{mix}} \geq wE_t^{\mathrm{L}}$ and $E_t^{\mathrm{mix}} \geq (1-w)E_t^{\mathrm{R}}$, giving the displayed inequality and regret at most $c(w)$. The constant is sharp: if $E_t^{\mathrm{L}}/E_t^{\mathrm{R}} \to \infty$ then $E_t^{\mathrm{mix}}/E_t^{\mathrm{L}} \to w$ and the regret approaches $-\log w$, and symmetrically for the robust expert. Hence no smaller uniform constant holds for that w, and $c(w)$ is minimized where $-\log w = -\log(1-w)$, uniquely at $w = 1/2$ with value $\log 2$. □

The advice-mismatch identity. Proposition 6 rests on an exact one-step identity for the expected log growth of a Kelly bettor with a possibly stale forecast.

Theorem 10 (Advice-mismatch growth identity)**.** *Consider one certification step with $Y \sim \mathrm{Ber}(p)$, $0 < p < m < 1$, and a ledger forecast $r \in (0, m)$. With the Kelly factor $B_r(Y) = 1 + \frac{m-r}{m(1-m)}(m - Y)$,*

$$\mathbb{E}_p[\log B_r(Y)] = \mathrm{kl}(p\|m) - \mathrm{kl}(p\|r).$$

The oracle forecast $r = p$ maximizes expected log growth, stale advice loses exactly $\mathrm{kl}(p\|r)$ relative to it, and ledger evidence grows in expectation if and only if $\mathrm{kl}(p\|r) < \mathrm{kl}(p\|m)$.

Proof of Theorem 10. With $\lambda = (m-r)/\{m(1-m)\}$, $B_r(1) = 1 - \lambda(1-m) = r/m$ and $B_r(0) = 1 + \lambda m = (1-r)/(1-m)$. Hence

$$\begin{aligned}\mathbb{E}_p \log B_r(Y) &= p \log \frac{r}{m} + (1-p) \log \frac{1-r}{1-m} \\ &= \left[p \log \frac{p}{m} + (1-p) \log \frac{1-p}{1-m}\right] - \left[p \log \frac{p}{r} + (1-p) \log \frac{1-p}{1-r}\right] = \mathrm{kl}(p\|m) - \mathrm{kl}(p\|r).\end{aligned}$$

The second term is nonnegative and zero only at $r = p$. The rejection direction follows by replacing (Y, p, m, r) with $(1-Y, 1-p, 1-m, 1-r)$. □

Proof of Corollary 8. Since $E_t^{\mathrm{mix}} \geq wE_t^{\mathrm{L}}$ and $E_t^{\mathrm{mix}} \geq (1-w)E_t^{\mathrm{R}}$ at every t, the mixture reaches A no later than the ledger expert reaches A/w or the robust expert reaches $A/(1-w)$. Taking first crossing times, with a missing crossing at time ∞, gives the claim. No growth-rate or overshoot assumption is used. □

Proof of Proposition 6. Write $f(r, y) = \log\{1 + \lambda(r)(m-y)\}$ with $\lambda(r) = [m-r]_+/\{m(1-m)\}$, so that $f(r, 1) = \log(\min\{r, m\}/m)$ and $f(r, 0) = \log\{(1 - \min\{r, m\})/(1-m)\}$. (i) The ledger forecast $r = p - \eta$ is fixed, so $t^{-1} \log E_t^{\mathrm{L}} = t^{-1} \sum_{s \leq t} f(r, Y_s) \to \mathbb{E}_p f(r, Y)$ by the strong law. If $r \geq m$ the stake is zero and $f \equiv 0$. If $0 < r < m$, Theorem 10 gives $\mathbb{E}_p f(r, Y) = \mathrm{kl}(p\|m) - \mathrm{kl}(p\|r)$. If $r \leq 0$ the clipped forecast is 0, $f(0, 1) = -\infty$, and the wealth is zero from the first unit loss, which occurs almost surely because $p > 0$. (ii) On $(0, p)$ the map $r \mapsto \mathrm{kl}(p\|r)$ is continuous and strictly decreasing from $+\infty$ to 0, so $r_\star$ exists and is unique; on $(p, 1)$ it is strictly increasing. Hence $\mathrm{kl}(p\|r) < \mathrm{kl}(p\|m)$ exactly for $r \in (r_\star, m)$, that is $\eta \in (\eta_-, \eta_+)$; $\mathrm{kl}(p\|r) = 0$ only at $r = p$; and $\mathrm{kl}(p\|r) > \mathrm{kl}(p\|m)$ for $r \in (0, r_\star)$, that is $\eta \in (\eta_+, p)$. (iii) The truncated robust forecast $\hat{p}_{t-1} = \max\{\varepsilon, \bar{Y}_{t-1}\}$ (with the proxy mean as the forecast at $t = 1$) converges to p almost surely because $p \geq \varepsilon$. On $[\varepsilon, 1]$, $f(\cdot, y)$ is continuous and bounded for $y \in \{0, 1\}$, because $f(r, 1) \geq \log(\varepsilon/m)$, so it is uniformly continuous with some modulus ω. Then $t^{-1} \sum_{s \leq t} |f(\hat{p}_{s-1}, Y_s) - f(p, Y_s)| \leq t^{-1} \sum_{s \leq t} \omega(|\hat{p}_{s-1} - p|) \to 0$ by Cesàro averaging, and $t^{-1} \sum_{s \leq t} f(p, Y_s) \to \mathbb{E}_p f(p, Y) = \mathrm{kl}(p\|m)$ by the strong law and Theorem 10 with $r = p$. (iv) Pathwise $\frac{1}{2} \max\{E_t^{\mathrm{L}}, E_t^{\mathrm{R}}\} \leq E_t^{\mathrm{mix}} \leq \max\{E_t^{\mathrm{L}}, E_t^{\mathrm{R}}\}$, so $t^{-1} \log E_t^{\mathrm{mix}}$ has the same limit as $t^{-1} \max\{\log E_t^{\mathrm{L}}, \log E_t^{\mathrm{R}}\}$, namely $\max\{G_L(\eta), \mathrm{kl}(p\|m)\}$, which equals $\mathrm{kl}(p\|m)$ because $G_L(\eta) \leq \mathrm{kl}(p\|m)$ by (ii). □

Proof of Corollary 9. For $\eta \in (\eta_-, \eta_+)$ the ledger log wealth $L_t = \sum_{s \leq t} f(r, Y_s)$ is a random walk with i.i.d. increments of mean $G_L(\eta) > 0$, whose largest increment is $f(r, 0) = c_\eta$. Let $T_{\mathrm{L}}(A')$ be its first passage above $\log A'$; it is finite with finite mean because the drift is positive and the increments are bounded. Before passage $L_t < \log A'$, so $L_{T_{\mathrm{L}} \wedge n} \leq \log A' + c_\eta$ for every n. Wald's identity (Wald, 1947) gives $G_L(\eta)\mathbb{E}[T_{\mathrm{L}} \wedge n] = \mathbb{E}[L_{T_{\mathrm{L}} \wedge n}] \leq \log A' + c_\eta$, and monotone convergence gives $\mathbb{E}[T_{\mathrm{L}}(A')] \leq (\log A' + c_\eta)/G_L(\eta)$. Corollary 8 with $w = 1/2$ gives $T_{\mathrm{port}}(K/\delta) \leq T_{\mathrm{L}}(2K/\delta)$ pathwise; taking expectations with $A' = 2K/\delta$ proves the first bound, and $A' = K/\delta$ gives the

ledger's. Their difference is $\log 2/G_L(\eta)$. At $\eta = 0$, $G_L(0) = \mathrm{kl}(p\|m)$. Finally, $\mathrm{kl}(1-\beta\|\delta) = (1-\beta)\log(1/\delta) + O(1)$ as $\delta \to 0$ and $\mathrm{kl}(p\|q_1) \to \mathrm{kl}(p\|m)$ as $q_1 \downarrow m$, which gives the stated comparison with Theorem 1. □

Stream-level extension. If predictable block budgets satisfy $\sum_b \delta_b \le \delta_{\mathrm{total}}$, conditional block validity and a union bound give stream-level false-certification probability at most δ_{total}. The reported experiments use block-level $\delta = 0.05$.

E A MATCHING CERTIFIER AND THE COST OF LABEL CAPS

This appendix states and proves the upper-bound side of Theorem 1 in the Bernoulli certification model of Section 4. For a queried label X and $p_q(x) = q^x(1-q)^{1-x}$, let $z(X) = \log\{p_{q_1}(X)/p_{q_0}(X)\}$ and let $Z_{i,n}$ be the sum of the first n increments at coordinate i. With $a = \log(M/\beta)$ and $b = \log(1/\delta)$, VIGIL-LR processes coordinates in deployment order and queries coordinate i until $Z_{i,n} \le -b$, when it certifies the coordinate and continues, or $Z_{i,n} \ge a$, when it enters permanent fallback.

Theorem 11 (Matching expected-label upper bound)**.** VIGIL-LR *satisfies Eq. 4 over the composite safe and unsafe classes, and* $\mathbb{E}_0[\sum_i N_i] \le M\{\log(1/\delta) + c_-\}/\mathrm{kl}(q_0\|q_1)$ *with* $c_- = \log\frac{1-q_0}{1-q_1}$. *Consequently, for* $q_0, q_1 \in [\eta, 1-\eta]$, $\beta \le \beta_0 < 1$, *and* $\delta \to 0$, *the expected-label minimax rate is* $\Theta\big(M(q_1-q_0)^{-2}\log(1/\delta)\big)$.

Proof. Honesty. Let $i^\star$ be the first unsafe coordinate reached while the procedure is active. At mean q_1, $\exp(-Z_{i^\star,n})$ is a nonnegative martingale with unit start, and at larger means a supermartingale by the monotone likelihood ratio of the Bernoulli family. Ville's inequality gives $\Pr(\inf_n Z_{i^\star,n} \le -b) \le \delta$. If $i^\star$ instead reaches the upper boundary, absorbing fallback prevents every later certificate, so no δ/M split is needed. *Liveness.* Under P_0, $\exp(Z_{i,n})$ is a nonnegative martingale, and a supermartingale at safer means, so Ville's inequality and a union bound give $P_0(\exists i, n : Z_{i,n} \ge a) \le Me^{-a} = \beta$; each walk has negative drift and reaches $-b$ almost surely unless it first reaches a. *Cost.* Under P_0 the increment mean is $-\mathrm{kl}(q_0\|q_1)$, and the lower crossing overshoots by at most c_-, so Wald's identity (Wald, 1947) on $\tau_i \wedge r$ gives $\mathrm{kl}(q_0\|q_1)\mathbb{E}_0[\tau_i \wedge r] \le b + c_-$; monotone convergence and summation give the bound. The upper boundary is reached only on the fallback event and contributes no $\log M$ term to expected safe-path cost. The rate follows by combining with Theorem 1, since $\mathrm{kl}(q_0\|q_1) = \Theta((q_1-q_0)^2)$ on $[\eta, 1-\eta]^2$ and $\mathrm{kl}(1-\beta\|\delta) = \log(1/\delta)(1+o(1))$. □

Proposition 12 (Deterministic caps cost more)**.** *Let* $\rho = 1 + (q_1-q_0)^2/\{q_0(1-q_0)\}$. *An auditor that queries at most* n *labels at each coordinate almost surely needs*

$$n \;\ge\; \max\left\{ \frac{\mathrm{kl}(1-\beta\|\delta)}{\mathrm{kl}(q_0\|q_1)},\; \frac{\mathrm{kl}(\delta\|1-\beta)}{\mathrm{kl}(q_1\|q_0)},\; \frac{\log\big(1+4M(1-\beta-\delta)^2\big)}{\log\rho} \right\},$$

and the fixed-sample rule VIGIL-CAP, *which certifies a coordinate when its* n*-sample mean is at most* $(q_0+q_1)/2$ *and otherwise falls back permanently, satisfies Eq. 4 whenever* $n \ge 2(q_1-q_0)^{-2}\max\{\log(M/\beta), \log(1/\delta)\}$. *For interior parameters the cap therefore costs* $\Theta\big((q_1-q_0)^{-2}\max\{\log M, \log(1/\beta), \log(1/\delta)\}\big)$ *per coordinate, a* $\log M$ *factor that adaptive stopping removes.*

Proof. The first term is Theorem 1 with $\mathbb{E}_0[N_i] \le n$. The second reverses the relative-entropy direction: $D_{\mathrm{KL}}(P_i^{\mathrm{tr}}\|P_0^{\mathrm{tr}}) \le n\,\mathrm{kl}(q_1\|q_0)$ and data processing through the all-certified event gives $n\,\mathrm{kl}(q_1\|q_0) \ge \mathrm{kl}(\delta\|1-\beta)$. For the third, let $L_i = dP_i^{\mathrm{tr}}/dP_0^{\mathrm{tr}}$ and $Q = M^{-1}\sum_i P_i$. For $i \ne j$ adaptive querying preserves $\mathbb{E}_0[L_iL_j] = 1$, and L_i^2/ρ^{N_i} is a nonnegative P_0-martingale with $N_i \le n$, so $\mathbb{E}_0[L_i^2] \le \rho^n$ and $\chi^2(Q\|P_0) \le (\rho^n - 1)/M$. The all-certified event has probability at least $1-\beta$ under P_0 and at most δ under Q, so $1-\beta-\delta \le \mathrm{TV}(P_0, Q) \le \frac{1}{2}\sqrt{(\rho^n-1)/M}$. For the upper bound, Hoeffding's inequality (Hoeffding, 1963) and a union bound over M coordinates give liveness, and the first unsafe coordinate reached passes its test with probability at most $\exp\{-n(q_1-q_0)^2/2\} \le \delta$, after which absorbing fallback prevents later certificates. □

F ADDITIONAL THEORETICAL REMARKS

This section collects remarks that the main text states briefly: the scope of the lower bound, why each assumption of the validity theorem is needed, and how to read the phase diagram.

Scope of Theorem 1. The expectation is taken under P_0, the stationary law on which residual transport is exactly true. The cost appears only because the auditor must also be honest over the class containing hidden single-coordinate violations, so it applies to any use of historical labels: prediction-powered correction, active sampling, shared ledgers, and randomized validation change the adaptive query rule inside the transcript, not the information available to it. For $\Delta = q_1 - q_0 \to 0$ the per-coordinate cost is $\Omega\big(q_0(1-q_0)\Delta^{-2}\,\mathrm{kl}(1-\beta\|\delta)\big)$, the residual-variance term one expects from a prediction-powered analysis, multiplied by M coordinates. A sequential likelihood-ratio certifier with absorbing fallback attains the bound up to an overshoot constant, closing the expected-label minimax rate at $\Theta\big(M(q_1-q_0)^{-2}\log(1/\delta)\big)$, and a deterministic per-coordinate cap pays an additional $\log M$ (Appendix E). The bound is stated for i.i.d. Bernoulli draws; it does not transfer verbatim to the finite without-replacement pools of Section 5, where complete enumeration resolves any candidate after N labels.

Necessity of the assumptions of Theorem 4. Each assumption is needed. If rows are not sampled uniformly, the conditional mean of the next loss need not exceed $m_t(k)$ and the betting factor can have mean above one. If a stake depends on the next row, for example on its proxy value, it is no longer predictable and the supermartingale property fails. If thresholds or advice depend on target-role outcomes, the null itself becomes data-dependent and the level is no longer controlled.

Reading the phase diagram. Proposition 6 says three things a reader can check against Figure 2. History can never improve the evidence *rate*: accurate advice only matches the rate the robust expert reaches by learning, so the value of good history is a finite-sample head start, which is exactly where the empirical frontiers separate (Figure 3). The damage from bad history is asymmetric: pessimistic stale advice makes the ledger stop betting and costs only power, while optimistic stale advice beyond η_+ makes it lose wealth, so the ledger's evidence decays to zero and it certifies only if it crosses the threshold early by chance. The portfolio's rate is the vigilance rate $\mathrm{kl}(p\|m)$ for every η, so it never inherits either failure. Corollary 9 makes the finite-sample price of not knowing η explicit.

Stale-advice regimes in the numerical evaluation. In the numerical evaluation of Figure 2, exact advice gives mean labels of 430.3 for the ledger expert, 492.3 for the portfolio, and 790.6 for the robust expert. At $\eta = -0.06$ the ledger never bets and does not certify within 6,000 labels on any path, while the portfolio needs 913.4. At $\eta = +0.06$ the ledger fails to certify within 6,000 labels on 38.7% of paths, while the portfolio needs 751.0, slightly below the robust expert's 784.1; the ledger still crosses on 61.3% of paths despite its negative drift, and by Corollary 8 the portfolio stops no later than the earlier of the two experts' crossings at level $2K/\delta$. This is the qualitative pattern of the bad-history stress test in Section 7.2.

G DATA AND COMPARATOR DETAILS

This section describes how the data roles are built and how the PPRM-style comparator maps to the original method.

Role construction. DICES-990 and CIFAR-10N use deterministic checksum role assignment with 40% training, 20% ledger, 10% calibration, and 30% final roles. Whole DICES conversations are assigned to one role, and every one of the 72,103 DICES ratings and 50,000 CIFAR-10N images is assigned. CIFAR-10N additionally draws proxy-training, ledger and calibration, and final outcomes from its three disjoint human annotation sets. The primary evaluation covers 18 final blocks, five seeds, nine budgets, and eight methods, and the external evaluation covers six blocks with the same seeds, budgets, and methods.

PPRM-style comparator. PPRM monitors running deployed risk under shift using prediction-powered residual correction and anytime-valid CM-EB bounds (Zhang et al., 2026). Our comparator

Table 5: External blocks at full budget. Both methods resolve every decision correctly without false certification. Labels are mean ± SD over five seeds; the ratio is of mean labels.

Block	Portfolio labels ↓	PPRM-style labels ↓	Label ratio ↓	Decision (desc.)
CIFAR-10, noise severity 0	$2{,}925.2 \pm 2{,}581.2$	$4{,}684.6 \pm 2{,}288.4$	0.624	release
CIFAR-10, noise severity 1	165.8 ± 66.6	354.4 ± 71.4	0.468	release
CIFAR-10, noise severity 2	81.6 ± 46.6	221.2 ± 15.1	0.369	release
CIFAR-10, noise severity 3	60.0 ± 0.0	202.6 ± 0.5	**0.296**	release
SVHN (out of distribution)	$5{,}752.4 \pm 2{,}671.8$	$8{,}981.6 \pm 2{,}821.8$	0.641	reject
Naval regression	305.4 ± 59.7	333.6 ± 13.2	0.916	release

keeps the estimator $\widehat{R}_n = \overline{Q}_{\text{pool}} + n^{-1} \sum_{t \le n} (Y_{\pi_t} - Q_{\pi_t})$ and a CM-EB-style sequential radius but targets the mean of a fixed finite pool sampled by the same uniform without-replacement prefix, with the same thresholds, seeds, candidate sharing, and budget checkpoints as the portfolio. It returns release, rejection, or abstention, so label counts are never compared at unequal decision coverage. We call it "PPRM-style" throughout and do not transfer the original PPRM theorem to this setting.

H PER-BLOCK AND EXTERNAL RESULTS

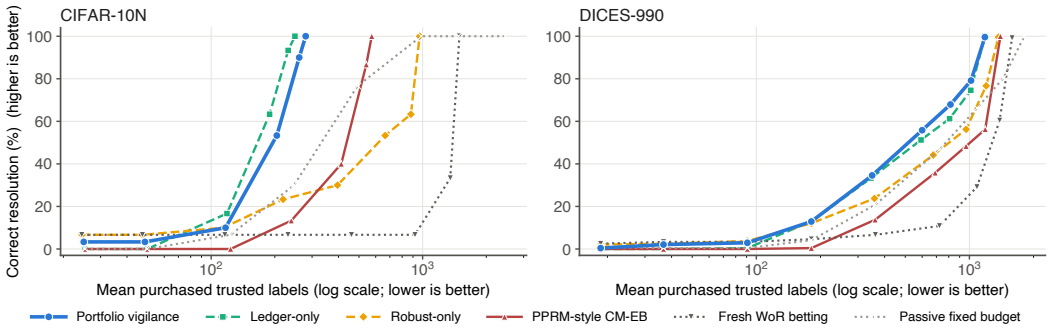


Figure 3: Correct resolution (↑) against mean purchased labels (↓, log scale) across the nine budget checkpoints, averaged over blocks and seeds.

Figure 4 shows the full-budget label ratio of every held-out block. The six U.S.-locale block ratios range from 0.422 to 0.849, while the six India-locale ratios range from 0.858 to 0.999. Across the 12 blocks the ratio has Spearman correlation -0.91 with the distance between block risk and threshold and $+0.72$ with the ledger forecast error: the U.S. blocks are further from the decision boundary and better predicted by history, which is where betting has the most to gain. These are descriptive associations at $n = 12$, not demographic inference. All six CIFAR-10N blocks favor the portfolio (ratios 0.275 to 0.582).

Table 5 applies the unchanged method to the six external blocks: four CIFAR-10 blocks at controlled Gaussian-noise severities 0 to 3 (10,000 rows each), SVHN as an out-of-distribution block (26,032 rows), and UCI Naval Propulsion (2,387 rows), each with a fixed candidate model and a cheap loss proxy. The gain is largest on the higher-severity CIFAR blocks, where either history or quickly learned target residuals give a decisive betting direction, and nearly vanishes on Naval, where both methods already resolve quickly. Budget checkpoints and seeds are correlated, so we do not convert the absence of false certifications into an independence-based error estimate; validity rests on Theorem 4.

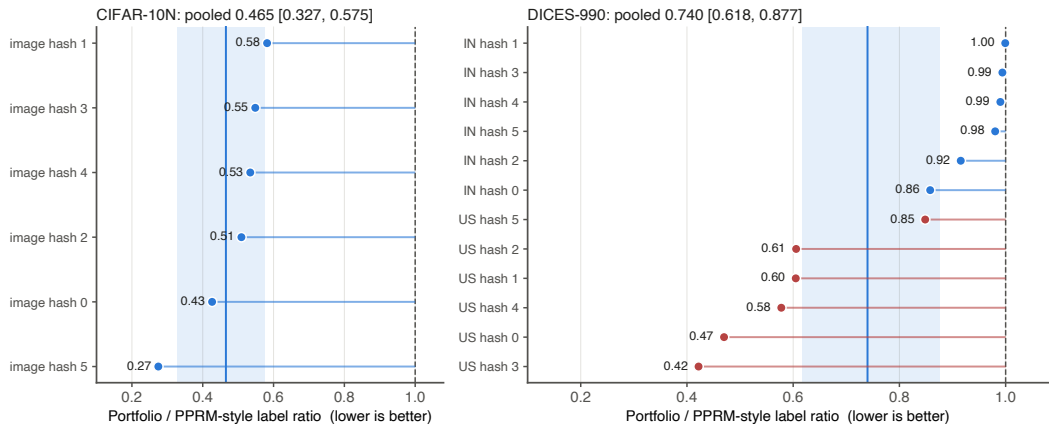


Figure 4: Per-block portfolio/PPRM-style label ratios (↓) at full budget, with the pooled ratio and its 95% interval (line and band) and parity (dashed). DICES-990: blue for India-locale, red for U.S.-locale blocks.

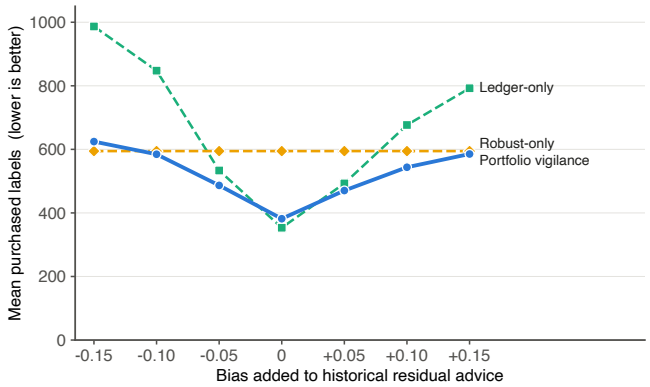


Figure 5: Mean purchased labels (↓) as the historical advice is biased, on synthetic finite pools.

I CORRUPTED ADVICE, MIXTURE WEIGHT, AND SCALING

These diagnostics evaluate portfolio vigilance on synthetic finite pools, separately from the natural-data analysis. Table 6 reports the advice-corruption grid and Table 7 the gap and error-budget scaling. Figure 5 plots the corruption grid.

Table 6: Mean purchased labels (↓) under biased advice ($N = 1{,}200$, threshold 0.20). Last column: portfolio labels over the cheaper component.

Advice bias	Ledger-only ↓	Portfolio ↓	Robust-only ↓	Portfolio / better component
−0.15	986.7	624.5	594.5	1.050
−0.10	847.7	584.6	594.5	0.983
−0.05	533.5	486.6	594.5	0.912
0	353.4	381.7	594.5	1.080
+0.05	492.8	470.4	594.5	0.955
+0.10	676.7	543.7	594.5	0.915
+0.15	792.3	585.4	594.5	0.985

Mixture weight. On the same construction with 500 replications, a safe pool (risk 0.20) with exact advice needs 464.5, 113.0, and 101.8 mean labels at ledger weight 0, 0.5, and 1; with misleading advice (forecast 0.5) it needs 464.5, 493.1, and 734.5. An unsafe pool (risk 0.40) needs 270.5, 110.6, and 98.2 with exact advice and 270.5, 308.5, and 695.5 with misleading advice (forecast 0.1). Empirical error rates are at most 0.016 in every cell. The weight trades efficiency between the two advice regimes and does not affect validity.

Risk gap. With safe risk 0.20, unsafe risk 0.40, and thresholds giving gaps from 0.02 to 0.15, mean labels with aligned advice fall from 679.4 to 61.0 on the safe side and from 672.8 to 48.4 on the unsafe side; misleading advice raises every cost but preserves the monotone pattern, so distance from the decision boundary remains the dominant difficulty.

Table 7: Portfolio scaling with aligned advice ($N = 5{,}000$, $\beta = 0.10$, threshold 0.30) against the lower bound of Theorem 1.

Gap	δ	Mean labels ↓	Lower bound	Ratio
0.03	0.05	1,145.9	281.9	4.06
0.05	0.05	468.0	102.4	4.57
0.075	0.05	220.1	45.9	4.80
0.10	0.05	121.9	26.0	4.69
0.15	0.05	54.1	11.6	4.67
0.10	0.01	188.8	41.7	4.52
0.10	0.10	91.0	19.2	4.74

J NATURAL-DATA SENSITIVITY

These analyses keep the roles, outcomes, sampling design, random orders, and certifier of the primary study and vary one factor at a time. Table 8 varies the threshold margin added to ledger risk.

Table 8: Threshold-margin sensitivity. Res@20: correct resolution at 20% budget (portfolio/PPRM). The portfolio makes no false certification at any margin.

Dataset	Margin	Ratio ↓	Portfolio labels ↓	PPRM labels ↓	Res@20 ↑
CIFAR-10N	0.020	0.428	938.9	2,032.7	0.100/0.000
CIFAR-10N	0.030	0.403	624.4	1,452.8	0.367/0.000
CIFAR-10N	0.050	0.465	279.9	574.6	0.900/0.400
CIFAR-10N	0.075	0.451	158.3	314.3	1.000/0.900
CIFAR-10N	0.100	0.414	98.4	211.6	1.000/1.000
DICES-990	0.020	0.729	1,367.1	1,782.2	0.192/0.046
DICES-990	0.030	0.674	1,245.8	1,707.9	0.225/0.054
DICES-990	0.050	0.740	1,182.2	1,398.3	0.346/0.138
DICES-990	0.075	0.709	1,048.7	1,241.9	0.513/0.275
DICES-990	0.100	0.649	860.5	1,089.5	0.642/0.458

CIFAR-10N observability. The primary proxy uses canonical class plus released worker and annotation-time side information. Retrained on canonical class only, the portfolio/PPRM ratio is 0.4658 (279.8 against 574.3 labels), essentially the primary 0.4654, with 90% against 40% resolution at 20% budget and zero portfolio false certifications.

A non-LLM evaluator-version change on DICES. We define an older cheap evaluator as a 10k-feature unigram TF-IDF logistic model and the current evaluator as the primary 30k-feature unigram-plus-bigram model, compute historical advice against the old evaluator, and expose the current evaluator on the target pool. At full budget the portfolio needs 1,184.8 labels with old-to-current advice against 1,182.2 with aligned advice, ledger-only 1,181.1 against 1,181.4, and robust-only 1,368.8 in both, with no false certification and the same single false rejection. This modest

change leaves the historical forecast informative, so it shows stability under one natural evaluator replacement rather than worst-case drift.

DICES block heterogeneity. Per-block ratios are 0.858, 0.999, 0.916, 0.994, 0.990, and 0.980 for India-locale hash blocks 0 to 5, and 0.470, 0.605, 0.606, 0.422, 0.578, and 0.849 for the U.S.-locale blocks. The ratio has Spearman correlation -0.91 with both the mean and the minimum absolute distance between block risk and threshold, and $+0.72$ with the mean absolute ledger forecast error.

K CONTROLLED LLM EVALUATOR-CONFIGURATION SHIFT

We design both LLM studies after fixing the primary protocol, so they complement rather than replace the DICES-990 and CIFAR-10N results. An initial single-run study judges the DICES ledger and final items once per rubric; the repeated-judge study reported in Section 7.3 judges every item three times per rubric and adds ToxicChat.

K.1 INITIAL SINGLE-RUN STUDY

Design and blinding. Both judges use the same model, GPT-5.6 Luna, with the same reasoning setting and a strict structured-output schema. Judge A uses a legacy narrow safety rubric and Judge B an updated contextual-risk rubric, and both return probabilities and NO/YES/UNSURE labels for harmful content, unfair bias, misinformation, and overall unsafe content. The judge-visible payload contains only an opaque record identifier, the conversation context, and the assistant response; human labels, role assignment, thresholds, historical outputs, certification state, and the research hypothesis are withheld. The deterministic DICES item split supplies 201 ledger and 298 final conversations, 499 per judge and 998 judge calls in total, joined back to 14,620 ledger and 21,702 final rating rows. The analysis reuses the primary random orders, budgets, δ, β, and ledger-risk-plus-0.05 thresholds. Judge A supplies the historical ledger and Judge B the current proxy in $A \to B$; $A \to A$ and $B \to B$ are stationary controls.

Provenance. All 998 judge calls complete successfully and report the intended model, and re-parsing every raw response reproduces the normalized judgments exactly. The two judges consume 659,890 input and 76,998 output tokens in total.

Shift diagnostics. On the 298 final conversations the mean absolute A/B probability difference is 0.0708 and the discrete A/B disagreement rate across the four outcomes is 8.98%. Residual mismatch for harmful content, unfair bias, misinformation, and overall unsafe content is $(0.0307, 0.0118, 0.0083, 0.0103)$ under $A \to A$, $(0.0162, 0.0474, 0.0579, 0.0873)$ under $A \to B$, and $(0.0243, 0.0006, 0.0142, 0.0003)$ under $B \to B$. For these diagnostics larger values indicate more evaluator change, not better performance. The update is not larger on every outcome, but it creates clear mismatch for unfair bias, misinformation, and especially overall safety.

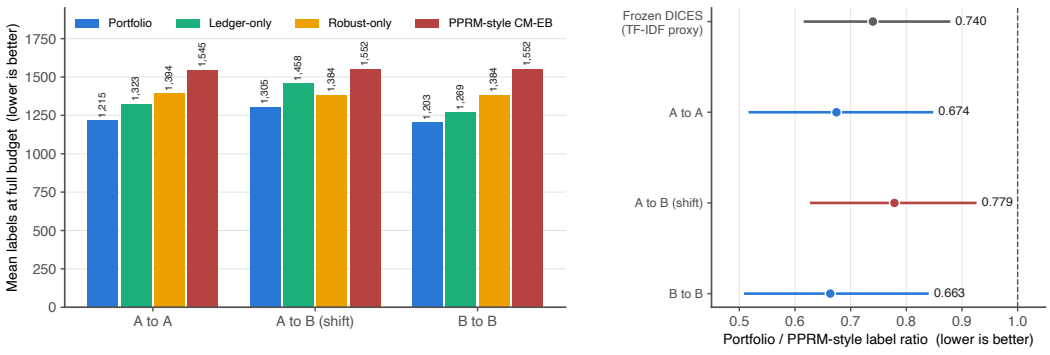


Figure 6: LLM evaluator-configuration study. Left: mean full-budget labels (↓). Right: portfolio/PPRM-style label ratio (↓) with 95% intervals, beside the primary DICES-990 result.

Table 9: LLM evaluator-configuration study at full budget. FC/FR over 240 decisions; P/P: portfolio/PPRM-style label ratio with its 95% interval.

Condition	Method	Labels ↓	Resolution ↑	FC ↓	FR ↓	AUC ↑	P/P ratio ↓
$A \to A$	Portfolio	1,214.9	0.9958	0	1	0.559	0.674 [0.519, 0.846]
	Ledger-only	1,323.4	1.0000	0	0	0.498	
	Robust-only	1,394.1	0.9958	0	1	0.505	
	PPRM-style CM-EB	1,545.1	1.0000	0	0	0.336	
	Fresh WoR betting	1,586.1	0.9917	1	1	0.316	
	Passive fixed budget	1,808.5	1.0000	0	0	0.521	
$A \to B$	Portfolio	1,304.5	0.9958	0	1	0.545	0.779 [0.629, 0.924]
	Ledger-only	1,458.3	1.0000	0	0	0.452	
	Robust-only	1,383.5	0.9958	0	1	0.491	
	PPRM-style CM-EB	1,551.8	1.0000	0	0	0.325	
	Fresh WoR betting	1,586.1	0.9917	1	1	0.316	
	Passive fixed budget	1,808.5	1.0000	0	0	0.521	
$B \to B$	Portfolio	1,203.5	0.9958	0	1	0.573	0.663 [0.510, 0.838]
	Ledger-only	1,268.8	1.0000	0	0	0.522	
	Robust-only	1,383.5	0.9958	0	1	0.491	
	PPRM-style CM-EB	1,551.8	1.0000	0	0	0.325	
	Fresh WoR betting	1,586.1	0.9917	1	1	0.316	
	Passive fixed budget	1,808.5	1.0000	0	0	0.521	

Paired contrasts. Table 9 gives full-budget results for every method, and Figure 6 compares the portfolio's label ratio across conditions with the primary result. Under $A \to B$, paired block-bootstrap label differences (portfolio minus comparator) are -153.8 against ledger-only (95% CI $[-288.8, -40.1]$), -79.0 against robust-only ($[-148.3, -19.2]$), and -247.2 against PPRM ($[-426.8, -101.1]$), and portfolio AUC exceeds PPRM by 0.2196 ($[0.1720, 0.2729]$). Relative to $A \to A$, the ledger-only expert needs 134.9 more labels under the shift ($[2.8, 291.1]$) and the portfolio 89.6 more ($[-2.7, 206.6]$); relative to $B \to B$ the increases are 189.5 ($[33.5, 366.9]$) and 101.1 ($[23.3, 204.1]$). Robust-only and PPRM are identical between $A \to B$ and $B \to B$ by construction, since neither uses A-residual advice.

Reading. The pattern is the one M2 predicts: stale advice costs the ledger expert efficiency without touching validity, and the mixture loses less because it retains the target-adaptive expert. The effect is modest. The portfolio's advantage over PPRM under the shift (0.779) is smaller than under either stationary control and than in the primary study (0.740), and every condition's interval overlaps the others. The portfolio's one false rejection occurs in the same U.S.-locale block under the same random order in all three conditions, so it does not arise from the rubric change. Fresh WoR betting's one false certification, in an India-locale block, likewise recurs in every condition.

Integrity. The study covers the complete design of three conditions, 12 blocks, five seeds, nine budgets, and six methods, and an independent check recomputes every reported summary from the per-block results.

K.2 Repeated-judge study

Design. The model, rubrics, schema, and blinding are those of the single-run study. Each rubric judges every item in three new independent runs (A2 to A4 and B2 to B4): all 990 DICES conversations and all 5,654 human-annotated ToxicChat-0124 examples. The single-run outputs are kept separate and are not reused as repetitions. ToxicChat is split by stable hash, without inspecting labels, into 2,245 proxy-training, 1,169 ledger, 547 calibration, and 1,693 final examples, and the final role forms eight hash blocks; DICES keeps its 12 final blocks. The monitors use the mean of the three runs of each rubric, the five primary random orders, and the nine budgets, and they add a without-replacement Hoeffding CS to the comparators. The written protocol for this study planned 50 random orders; we ran the five primary orders. Orders are averaged within each block before

10,000 block-bootstrap replicates, so the reduced number of orders affects Monte Carlo stability but not the inferential unit. An automated audit confirms the complete design of 51,840 DICES and 34,560 ToxicChat method rows and the seed list.

Rubric change against run-to-run variation. Table 10 compares judge scores between runs of the same rubric with scores between rubrics, using 10,000 item-bootstrap replicates. The between-rubric change exceeds the within-rubric change in every role of both datasets; the table shows the final roles.

Table 10: Judge run-to-run variation against rubric change on final items: mean absolute score change and binary-label disagreement within rubric A, within rubric B, and between rubrics, and the excess of the between-rubric value over the within-rubric value with its 95% interval.

Dataset	Metric	Within A	Within B	Between	Excess
DICES-990 ($n = 298$)	Score change	0.0181	0.0365	0.0749	0.0476 [0.0331, 0.0633]
	Label disagreement	0.0162	0.0336	0.0910	0.0661 [0.0474, 0.0864]
ToxicChat ($n = 1{,}693$)	Score change	0.0139	0.0136	0.0516	0.0379 [0.0301, 0.0460]
	Label disagreement	0.0110	0.0122	0.0566	0.0450 [0.0357, 0.0546]

Full results. Table 11 gives full-budget label costs in all three conditions. The pessimistic cost charges the full block for any false certification, false rejection, or unresolved decision. The portfolio makes no false certification in any condition; on DICES-990 it makes one false rejection in each condition, as does robust-only. Moving from $B \to B$ to $A \to B$ raises the ledger-only cost from 1,274.7 to 1,445.9 labels on DICES-990 and from 130.6 to 150.2 on ToxicChat, while the portfolio rises from 1,214.3 to 1,305.2 and from 130.7 to 137.6.

Table 11: Repeated-judge study at full budget: mean labels (↓) and pessimistic labels (↓) over blocks and the five random orders. FC/FR: false certifications/rejections over all runs. Robust-only, PPRM, and the fresh and passive methods do not use A-residual advice, so their $A \to B$ and $B \to B$ values coincide.

		DICES-990			ToxicChat		
Condition	Method	Labels	Pessim.	FC/FR	Labels	Pessim.	FC/FR
$A \to A$	Portfolio	1,201.1	1,220.8	0/1	131.5	131.5	0/0
	Ledger-only	1,282.8	1,282.8	0/0	129.0	129.0	0/0
	Robust-only	1,369.2	1,389.0	0/1	129.5	134.3	0/1
	PPRM-style CM-EB	1,536.9	1,536.9	0/0	210.7	210.7	0/0
$A \to B$	Portfolio	1,305.2	1,325.1	0/1	137.6	137.6	0/0
	Ledger-only	1,445.9	1,445.9	0/0	150.2	150.2	0/0
	Robust-only	1,414.2	1,434.4	0/1	137.5	137.5	0/0
	PPRM-style CM-EB	1,544.3	1,544.3	0/0	208.5	208.5	0/0
$B \to B$	Portfolio	1,214.3	1,234.3	0/1	130.7	130.7	0/0
	Ledger-only	1,274.7	1,274.7	0/0	130.6	130.6	0/0
	Robust-only	1,414.2	1,434.4	0/1	137.5	137.5	0/0
	PPRM-style CM-EB	1,544.3	1,544.3	0/0	208.5	208.5	0/0
All	Fresh WoR betting	1,586.1	1,595.7	1/1	173.3	178.1	0/1
	WoR Hoeffding CS	1,808.5	1,808.5	0/0	211.6	211.6	0/0
	Passive fixed budget	1,808.5	1,808.5	0/0	211.6	211.6	0/0

Paired comparisons under $A \to B$. Table 12 gives portfolio-minus-comparator label differences with block-bootstrap intervals, the number of the nine pairings of one A run with one B run in which the portfolio is cheaper, and the range of the difference when any one of the five random orders is left out. Every comparison favors the portfolio except robust-only on ToxicChat, where the interval, the pairing count, and the leave-one-order-out range all include no difference. Pessimistic costs give the

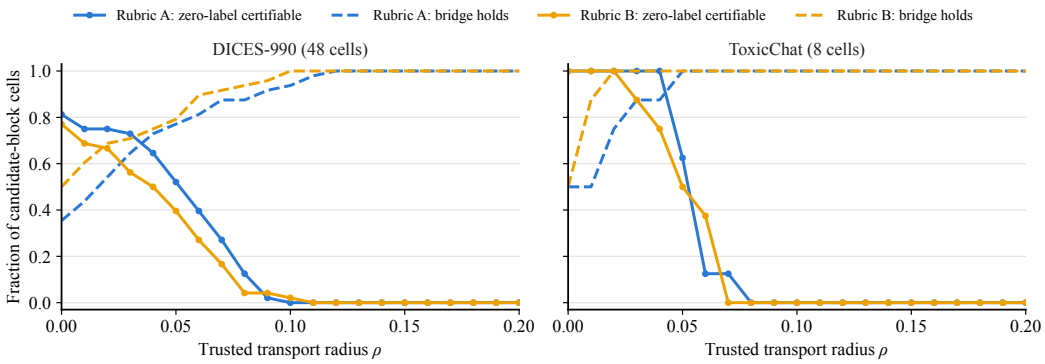


Figure 7: Trusted-transport radius ρ against the fraction of candidate-block cells that a zero-label certificate would certify (solid) and the fraction for which the bridge $\mu_E^{\text{tar}} \le \mu_E^{\text{hist}} + \rho$ holds on the realized target data (dashed), under each rubric's stationary judge grid of the repeated-judge study. The certificate uses the historical mean residual in place of the upper bound U_H, which makes it optimistic. A radius wide enough to hold removes most or all zero-label certificates on DICES-990.

same conclusions: on DICES-990 the differences are -120.8 ($[-218.0, -40.8]$) against ledger-only, -109.3 ($[-199.4, -32.9]$) against robust-only, and -219.2 ($[-349.7, -111.8]$) against PPRM, and on ToxicChat they equal the ordinary differences.

Table 12: Portfolio minus comparator, full-budget labels under $A \to B$ (negative favors the portfolio). CI: block-bootstrap 95% interval after averaging the five random orders. Pairings: A/B run pairings (of nine) in which the portfolio is cheaper. LOO: range over leaving out one random order.

Dataset	Comparator	Difference	95% CI	Pairings	LOO range
DICES-990	Ledger-only	-140.7	$[-256.2, -46.2]$	9/9	$[-148.8, -132.1]$
	Robust-only	-109.1	$[-199.3, -32.6]$	9/9	$[-118.6, -97.2]$
	PPRM-style CM-EB	-239.1	$[-394.4, -114.2]$	9/9	$[-272.4, -219.9]$
	WoR Hoeffding CS	-503.3		9/9	$[-520.4, -488.7]$
ToxicChat	Ledger-only	-12.7	$[-20.8, -5.9]$	9/9	$[-14.9, -10.3]$
	Robust-only	$+0.05$	$[-4.65, 4.05]$	4/9	$[-3.22, 2.75]$
	PPRM-style CM-EB	-70.9	$[-97.1, -42.6]$	9/9	$[-74.2, -68.3]$
	WoR Hoeffding CS	-74.1		9/9	$[-77.5, -69.9]$

The price of trusted transport. The same data illustrate M1. Theorem 3 says the zero-label certifiable set shrinks as the declared radius ρ grows; Figure 7 sets this against the fraction of cells for which the bridge happens to hold on the realized target data, a fact that only the target labels transport was meant to save can reveal. At $\rho = 0$ a zero-label certificate would cover 81% and 77% of the DICES-990 candidate-block cells under rubrics A and B, although the bridge holds for only 35% and 50%; every certificate issued where the bridge fails is unjustified. On DICES-990 the radius at which the bridge first holds everywhere (0.12 and 0.10) leaves 0% and 2% of the cells certifiable, so a bridge wide enough to be true removes the zero-label advantage. ToxicChat, with a single outcome and eight blocks, is more forgiving: a radius of 0.05 (rubric A) or 0.02 (rubric B) holds on every block and still certifies 62.5% and 100% of them, but nothing in the cheap data identifies that radius in advance (Proposition 2).

Endpoint calibration comparison. Noisy-but-Valid (Feng et al., 2026) tests a fixed endpoint rather than monitoring sequentially, so we report it only as a diagnostic. With its calibration-set estimates of the judge's true- and false-positive rates, it certifies none of the ten rubric-by-endpoint cases on ToxicChat and on the DICES conversation-level majority endpoints, including the eight

whose actual risk is below the threshold, so it makes no false certification but also no correct certification.

L HIDDEN FAILURE OF UNVERIFIED TRANSPORT

This study uses an earlier certifier family and different data from the rest of the paper, so we do not pool its results with Section 7; it asks only whether trusting a calibrated bridge without verifying it fails on real data. The data are ACSIncome and ACSPublicCoverage (Ding et al., 2021) for four states with year-separated roles (2014 training and ledger, 2015 bridge calibration, 2016 to 2018 held-out tests), logistic candidates, and a 60-cell descriptor partition that defines the bridge. TRANSPORT-FIRST releases every candidate the calibrated bridge covers and never re-checks a released candidate with labels; VIGIL-CAP is a fixed-sample variant that also releases bridge-covered candidates; the exploratory VIGIL-CELL verifies each bridge-covered cell with a sequential test on fresh labels. The shared-ledger oracle uses the bridge with zero labels, and the proxy-only control trusts the cheap scores alone.

For every held-out block, an injected copy raises the trusted losses of one proxy-score decile cell until every candidate exceeds its threshold by 0.02, leaving proxies, descriptors, and the bridge untouched: the real-data analogue of the hidden alternatives in Section 4. At the prespecified release margin of 0.02 the calibrated bridge (width 0.048) is wider than the margin, so it transports nothing and no method false-certifies. The larger margins, the calibrated 0.038 and 0.05, are exploratory and were chosen after this diagnostic.

Table 13: Hidden-failure stress test on Folktables: false certification on injected blocks (120 per margin) and fresh labels on paired clean blocks. Margins other than 0.02 are exploratory.

	Margin 0.02		Margin 0.038		Margin 0.05	
Method	FC ↓	Labels ↓	FC ↓	Labels ↓	FC ↓	Labels ↓
TRANSPORT-FIRST	0.000	$48{,}275$	0.417	$29{,}525$	$\mathbf{0.750}$	$10{,}816$
VIGIL-CAP	0.000	$20{,}472$	0.417	$12{,}795$	$\mathbf{0.750}$	$5{,}118$
VIGIL-CELL (exploratory)	0.000	$33{,}552$	0.000	$22{,}838$	0.000	$11{,}933$
Selective residual CS	0.000	$7{,}924$	0.000	$6{,}195$	0.000	$4{,}733$
PPRM-reset	0.000	$6{,}813$	0.000	$4{,}645$	0.000	$3{,}362$
Betting monitor	0.000	$18{,}749$	0.000	$8{,}695$	0.000	$10{,}644$
Fresh-label Hoeffding	0.000	$35{,}768$	0.000	$23{,}679$	0.000	$16{,}140$
Shared-ledger oracle	0.000	0	0.417	0	$\mathbf{0.750}$	0
Proxy-only control	1.000	0	1.000	0	1.000	0

Wherever the bridge transports, TRANSPORT-FIRST is block for block identical to the no-label oracle, false-certifying 0.417 and 0.750 of the injected blocks, while every fresh-label baseline catches all of them (Figure 8). The exploratory VIGIL-CELL stays at zero false certification on all 120 injected blocks at every margin; at margin 0.05, where the bridge covers 90 of 120 blocks, it uses $1{,}489 \pm 96$ labels on covered blocks against $1{,}685 \pm 126$ for PPRM and rejects an injected failure with a median of 346 labels against $7{,}175$. Aggregated over all blocks PPRM remains cheaper ($3{,}362$ against $11{,}933$), and because margin 0.05 was chosen after the diagnostics, this comparison is post hoc. These failures of blind transport motivate treating untrusted history as advice rather than as part of validity.

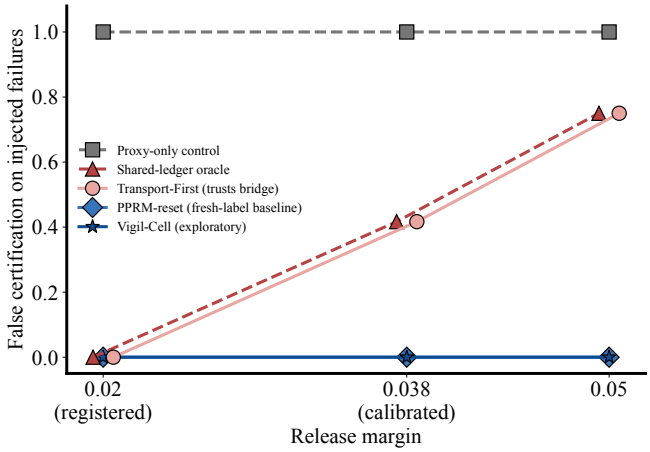


Figure 8: False certification on injected failures across the three tested margins (Table 13). The shared-ledger oracle and TRANSPORT-FIRST, which trusts the bridge without verifying it, rise together once the bridge transports; fresh-label baselines and the exploratory cell-verifying certifier stay at zero.