AUDITING MODEL FUSION UNDER ASYMMETRIC LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Systems that fuse several models under an asymmetric operational loss usually summarise each model by a *scalar*: a hard vote, or a reliability weight in place of a likelihood. We ask three questions about that practice, and answer each with a finite-sample object. *Is it the scalar that hurts?* A matched control fuses hard labels through one code path and varies only how each model is described: its fitted confusion matrix, against the symmetric channel carrying that matrix's own accuracy. This separates error structure from the change of representation that confounded earlier comparisons. The scalar description loses on all 10 seeds of the primary domain, by a paired $+106.360$ risk-weighted loss; the secondary domains show smaller or more sample-sensitive effects. *Can the evaluation tell?* An evaluation can be quantised more coarsely than the effect it measures. For a separate, smaller contrast against weighted averaging, we distinguish observed ranking sensitivity from uncertainty using an exact leave-one-out diagnostic—$11.1\times$ tighter than the bound it replaces—and a retrospective empirical-Bernstein calculation. *What can a deployed rule certify?* We give a one-sided finite-sample certificate that controls, at level δ, the probability of falsely declaring a deployed rule worse than a fixed comparator, built on where in the loss matrix the rule sends the rare class, and costing only the calibration split rather than the rare rows alone. It fires on all 10 seeds of the decisive domain under the *weaker* of two concentration bounds. Where it cannot fire it abstains, and we report the abstention.

1 INTRODUCTION

Decision systems in intrusion detection, predictive maintenance and fraud screening commit actions under a strongly asymmetric loss: missing a rare failure costs 1000 against 10 for over-escalating on a normal input, a 100:1 ratio (Table 3). Under such a loss the cost of an error is set by *where* the committed action lands, not by how often the system is wrong. Yet the fusion rules deployed in practice compress each model to a scalar before that decision is taken—a hard vote, or a reliability weight substituted for the likelihood—and a scalar cannot express where a model's errors go.

The symptom is a reversal. Among rules that see only hard labels or a scalar accuracy, which rule is best flips across domains: risk-weighted loss (RWL) swings $13.7\times$ on NSL-KDD (network intrusion), where voting fails catastrophically on the rare class, and $9.9\times$ the other way on AI4I (machine failure; Fig. 1a and Table 4). Giving the fuser each model's full class-conditional error structure removes most of it—established in earlier work on this system, which left two questions open that this paper is about.

The confound. The arm that removes the scalar-summary reversal changes two things at once. It replaces a scalar reliability with a fitted class-conditional channel, *and* it consumes hard labels rather than soft probability vectors. A reader is therefore entitled to attribute the repair to the change of representation rather than to error structure, and nothing in the earlier comparison excludes that. §4 removes the confound with a matched control that holds the representation fixed and varies only the channel, and the scalar arm loses on every seed.

The resolution limit. A rare-class row on which the system commits IGNORE is charged $L_{\max}$, so it moves RWL by $L_{\max}/n_{\text{test}}$ *whatever* the prevalence—the ceiling on what one row can do, since

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.

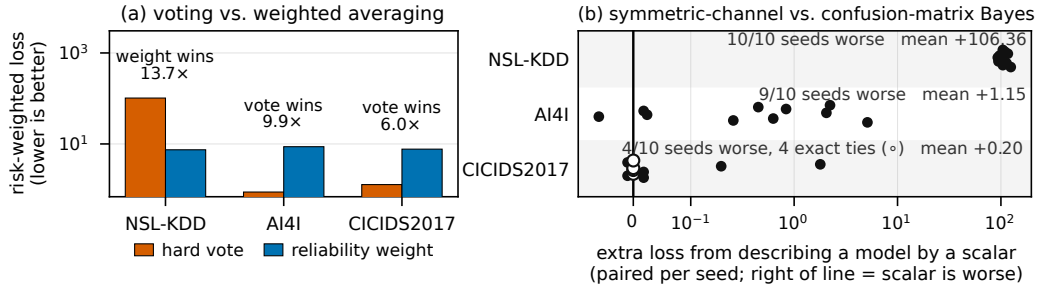

Figure 1: **The symptom, and the test this paper runs on it. (a)** Among rules that see only hard labels or a single accuracy, which rule is best reverses across domains, so a scalar rule chosen on one domain does not carry to another. **(b)** Holding the fusion code path and the label representation fixed and varying only how each model is described, the scalar arm is worse on every NSL-KDD seed—but on CICIDS2017 the two arms commit identical actions on four seeds, leaving almost no contrast for the evaluation to resolve.

a milder misstep on the same row costs less. On a domain whose evaluation split carries a median of four rare rows, that single row is worth $68.7\times$ the effect being measured. A non-significant result there is not evidence of a tie; it is a comparison that cannot adjudicate. But a ceiling is a bound, and §5 replaces it with two exact objects: the leave-one-out sensitivity ρ_{LOO}, which answers the question the ceiling only bounds, and the empirical-Bernstein sample size at which the comparison becomes resolvable.

What a deployed rule can certify. For rules already in production the useful statement is one-sided: reject, or say nothing. The population version is immediate from a lower bound on the deployed rule's rare-class risk. What ships, however, substitutes point estimates on both sides, and a point estimate cannot stand in for an upper bound. §6 closes that gap with a genuine finite-sample certificate.

Contributions. **(i)** A *matched* falsification (§4): one fusion code path, two channels, everything else held fixed, with the reading of both outcomes fixed in advance. **(ii)** A finite-sample account of *evaluation resolution* (§5): the exact one-row sensitivity ρ_{LOO} and the paired-risk sample size required to resolve an effect, which together separate "no effect" from "no resolution"—a distinction an aggregate significance test does not make, because a non-significant result is consistent with either. **(iii)** A one-sided finite-sample rejection certificate for deployed rules (§6), reported under the weaker of two concentration bounds, together with the domains on which it abstains. The hard-label sufficiency result of §3 is included as structural background for what the experiments vary, not as a contribution.

2 SETTING

Rounds, actions and loss. At each round the system observes features x_t and commits an action before a per-round deadline, from

$$\mathcal{A} = \{\textsc{ignore}, \textsc{monitor}, \textsc{escalate}, \textsc{emergency}\},$$

against a latent state $y_t \in \mathcal{Y}$,

$$\mathcal{Y} = \{\textsc{normal}, \textsc{degraded}, \textsc{critical}, \textsc{failure}\},$$

with $y^* = \textsc{failure}$ the rare class. The asymmetric loss $L : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ (Table 3) charges each of the four actions differently on a FAILURE row, IGNORE worst at 1000; given a belief b the committed action is $\arg\min_a \sum_y b_y L(a, y)$, and the reported metric is risk-weighted loss $\mathrm{RWL} = \frac{1}{T}\sum_t L(a_t, y_t)$. Two constants recur. $L_{\max} = 1000$ is the largest entry of L. $D_{\max} = \max_y[\max_a L(a, y) - \min_a L(a, y)] = 998.5$ is the largest loss *swing* available on a single row, and it, not $L_{\max}$, is the constant a *paired* comparison of two rules can move by.

Models, budget and orchestration. Each domain gets five trained scikit-learn models (logistic regression, k-NN, random forest, XGBoost, a soft-voting ensemble) with costs 0.01–5.0; on NSL-KDD the defining property is poor rare-class coverage (FAILURE accuracy 3.0%–27%). A fixed budgeted controller queries the cheap models first and escalates when oracle disagreement, the risk gap or low confidence crosses a threshold, under a fixed query budget and the per-round deadline. The controller is the *setting*, not a contribution: every comparison holds it fixed and varies only the fusion step, and the headline comparison is repeated under three other orchestrators in the supplement.

Data and protocol. NSL-KDD, AI4I 2020 and CICIDS2017 (a second, larger intrusion corpus) at deployment prevalence, mapped to the four-class severity scheme above. Two parameter paths run in this paper and they must not be conflated. The *fusion* arms of §4 estimate calibration quantities—temperatures, confusion matrices, per-model reliabilities and the class prior—on one held-out split, disjoint from evaluation (n_{test}=500, 10 seeds). The *certificate* path of §6 estimates none of them: the rule it certifies and the comparator it certifies against are fixed aggregators whose every parameter is set at construction and never fitted, so no quantity on that path is estimated from the rows used to certify (App. G). Within a seed every rule shares the model pool, the evaluation subset, the data order and the random stream, so paired comparisons are exact. Cross-seed significance uses an *exact* two-sided sign-flip permutation test on paired per-seed differences, whose p-value floor at n=10 is $2/2^{10}$=0.002.

Reproduction is asserted, not assumed. Every re-run checks that it reproduces the frozen run *bitwise* per seed and arm before its output is used; a seed that fails is excluded and reported as excluded. The uncertainty audit's cached replay is the exception: it reproduces $8/10$ AI4I seeds, so it supplies interval widths only. The precondition is stricter than it looks. Fixing the thread count is necessary but not sufficient—in our environment, runs completed under load failed bitwise reproduction at identical thread count and affinity—so an accepted re-run requires a quiescent machine, with wall-clock used only as an early-stopping heuristic and never as the acceptance criterion. App. C gives the measurements and says what they do and do not identify.

3 SCALAR RELIABILITY AS A CHANNEL MODEL

Background for what §4 varies; we claim no novelty for it, and the statements, proofs and counterexample are in App. E. Model one source as a channel $Q_{yc} = \Pr[Z{=}c \mid Y{=}y]$ over hard labels; summarising it by one accuracy α asserts the symmetric channel

$$S_\alpha(y, c) = \alpha \mathbb{1}\{c{=}y\} + \tfrac{1-\alpha}{C-1} \mathbb{1}\{c \neq y\}. \tag{1}$$

That is an identity, not an approximation: matching Q's posteriors at every observed label forces $Q = S_\alpha$ whenever $\alpha \neq 1/C$ (Thm. 6), *per model* and not across a product of K. Two consequences bound any scalar rule: equal accuracy does not determine the Bayes action (Cor. 7), and misspecification is free until it crosses a loss-induced margin whose radius shrinks as the loss sharpens (Lem. 8)—which is why the asymmetry is not incidental.

4 A MATCHED FALSIFICATION

A scalar reliability is a symmetric-channel assumption (Thm. 6). The empirical claim is that dropping that assumption is what repairs the scalar-summary reversal. The arm that established it does not show this, because it also changes the representation: it reads hard labels where the deployed rules read soft vectors. This section removes that confound.

Design. Both arms fuse hard labels identically,

$$P(Y{=}y \mid \hat{y}_{1:K}) \;\propto\; \pi(y) \prod_{k=1}^{K} P_k(\hat{y}_k \mid Y{=}y), \tag{2}$$

through one function on the same rows, the same calibration split, the same class prior and the same add-one Laplace smoothing. They differ in P_k and in nothing else. **Arm B** uses the fitted

confusion matrix Q_k, an estimate of $\Pr[\hat{y}_k{=}c \mid Y{=}y]$ on the held-out split—*supervised* confusion-matrix Bayes, Dawid–Skene-*style* (Dawid & Skene, 1979), since the split carries gold labels and no unsupervised EM is involved; in the classifier-fusion taxonomy it is the *naive Bayes combiner* (Kuncheva, 2014), a name we adopt rather than reintroduce the rule (App. D). **Arm A** uses $S_{\hat{\alpha}_k}$ from (1), the symmetric channel at model k's own accuracy. The estimator for $\hat{\alpha}_k$ was pinned before the arm was implemented, and it is deliberately the scalar summary of the very table Arm B uses: with N_k the Laplace-smoothed count table behind Q_k,

$$\hat{\alpha}_k = \operatorname{tr}(N_k) \big/ \textstyle\sum_{y,c} N_k(y,c). \tag{3}$$

Arm A is therefore exactly Arm B with each model restricted to reporting one number, and any difference is attributable to the channel's class-conditional structure rather than to the fusion arithmetic, the smoothing, or the representation. It does *not* isolate the off-diagonal destinations on their own. Replacing Q_k by $S_{\hat{\alpha}_k}$ constrains two things at once: it forces the diagonal $Q_k(y,y)$ to be constant in y, and it forces the errors to spread uniformly over the remaining classes. A difference therefore identifies the pair, not either member. Separating them would need a third channel that keeps each class's own diagonal competence $Q_k(y,y)$ and uniformises only the error destinations; the protocol admits no third hard-label arm, so we do not claim that separation.

Reading, fixed in advance. If the arms agree, the repair is explained by moving to a hard-label channel model at all, and class-conditional structure beyond a single global accuracy is *not* operative: agreement would mean the symmetric channel was adequate on these domains (Thm. 6), and the empirical claim narrows accordingly. If Arm A is worse, the full class-conditional channel does work that the tested symmetric-channel baseline does not reproduce—the reading this line of work has so far declined to assert. Both outcomes are reported as the headline of this section, and no third hard-label variant may be added after the fact. The primary endpoint is NSL-KDD; the other two domains are read through §5 and never as a "three-domain winner".

Result. Fidelity first, since it is what licenses reading the line as a matched comparison at all. Arm B re-runs the frozen protocol rather than reading its recorded value, and it reproduced that value bitwise on all 10 seeds (per-seed detail in Table 5). A seed that does not reproduce is excluded and reported as excluded, never averaged in; earlier attempts at this experiment did lose seeds that way and were discarded whole rather than repaired, so the count is printed to make "the only difference is the channel" a checked statement rather than an intended one.

The outcome is the second of the two readings fixed above: Arm A is worse on *every* seed, 112.295 against 5.935 RWL, a paired $\Delta = +106.360$ at the exact sign-flip floor (Fig. 1b), $p = 0.002$: restricting each model to one accuracy number costs more than an order of magnitude. The repair reported in earlier work is therefore not explained by the move to hard labels—class-conditional structure beyond a single global scalar is doing the work, the reading this line of work had so far declined to assert. Two limits carry over: the comparison identifies the pair of constraints $S_{\hat{\alpha}_k}$ imposes rather than the off-diagonal destinations alone, and it rests on sign consistency across seeds, not on any one seed's interval. The contrast §5 analyses is a different and far smaller one, so its sample sizes do not transfer here.

The two secondary domains. The protocol fixed NSL-KDD as the primary endpoint and required the others to be read through §5, not as two further votes; Table 6 does that per seed, and also prices the primary result more sharply than sign consistency can ($\rho_{\mathrm{LOO}} \le 0.0214$ on every seed, the arms differing on 284–316 of 500 rows). AI4I agrees in sign on 9 of 10 seeds, 2 of them one row from flipping. On CICIDS2017 the arms agree *exactly* on 4 seeds and differ on at most 2 rows anywhere, so $\Delta = +0.200$ on $4/10$ is not two channels tying but a domain with almost no contrast to resolve—reading that as a null is what §5 refuses.

5 WHEN CAN THE EVALUATION ADJUDICATE?

The loss is graded, not binary: on a FAILURE row IGNORE costs 1000, MONITOR 100 and ESCALATE 10. A *maximally costly* rare-class error—IGNORE on a FAILURE row—therefore moves RWL by $L_{\max}/n_{\mathrm{test}} = 2.0$ whatever the prevalence, and that is the *ceiling* on what any one row can do. At NSL-KDD's median 64 rare evaluation rows that ceiling is $1.3\times$ the measured effect;

at CICIDS2017's 27 it is $12.4\times$; at AI4I's median of *four* it is $68.7\times$. Tracing the outlying seeds shows the ceiling is not hypothetical: on two AI4I seeds the rule commits IGNORE on precisely *one* rare row out of three and four present, and that single row contributes 2.0 of roughly 2.5 RWL.

A ceiling, though, is a bound on the worst row the loss matrix admits, not a statement about the rows the experiment actually produced. Two exact objects replace it.

5.1 Exact One-Row Sensitivity

Let $d_i = L_{\mathrm{B}}(i) - L_{\mathrm{A}}(i)$ be the paired per-row loss difference between two rules on the evaluation split and $\Delta = \frac{1}{n}\sum_i d_i$. Deleting row i gives $\Delta_{-i} = (n\Delta - d_i)/(n-1)$, so the empirical winner reverses under a one-row deletion iff some d_i lies beyond $n\Delta$ on Δ's side. Define

$$\rho_{\mathrm{LOO}} = \frac{\max\{|d_i| \ : \ d_i\Delta > 0\}}{n\,|\Delta|}, \tag{4}$$

so that $\rho_{\mathrm{LOO}} > 1$ iff some single evaluation row can be deleted to reverse the empirical winner, and $\rho_{\mathrm{LOO}} \le 1$ certifies that none can. This is an exact finite-sample statement about the observed sample: no distributional assumption, no bound. It replaces the heuristic ratio on two counts—the correct worst-case constant for a *paired* comparison is $D_{\max}$ rather than $L_{\max}$, and the exact question is answerable rather than boundable.

The gap between bound and measurement is large, and it changes the verdict in both directions.

On NSL-KDD the worst-case ratio $D_{\max}/(n|\widehat{\Delta}|)$ exceeds 1 on 8 of 10 seeds—read literally, it declares eight seeds vulnerable to a single row. The exact statistic says the opposite: $\rho_{\mathrm{LOO}} \le 0.26$ on every seed and *no* seed has a row whose deletion flips it, a margin of nearly $4\times$ on the worst. The two disagree by a constant factor of 11.1, for a reason visible in the data: the arms differ by at most 90.0 on any single row, against the $D_{\max}$=998.5 the matrix admits. The bound prices a row the experiment never produces, and it does so against our own earlier reporting.

On AI4I the exact statistic instead *confirms* the concern, and localises it. Seeds 3 and 6 have $\rho_{\mathrm{LOO}} = 1.29$ and 1.20 with exactly one flipping row each, and in both cases that row is a rare-class row on which the two arms differ by 990.0—the swing from ESCALATE to IGNORE on a FAILURE row—the largest swing those two actions admit—realised. Every other AI4I seed has $\rho_{\mathrm{LOO}} \le 0.18$ and no flipping row. So the sign flip in AI4I's aggregate is not diffuse disagreement: it is two rows, identified, on two seeds, against eight seeds favouring the confusion arm. Reporting that as a tie would have been a statement about two rows.

CICIDS2017 sits between the two, and does so in a way worth reporting carefully. Its ceiling ratio is $12.4\times$ and its bound above 5 on every seed, while $\rho_{\mathrm{LOO}} \le 0.10$ with no flipping row on 9 of the 10 seeds: there, no *single* row decides the comparison, and the under-resolution is a statement about the aggregate effect relative to sampling variability rather than about one-row leverage—a distinction the ceiling column alone cannot draw. The tenth seed is the exception, and we name it because an earlier partial sweep of this domain contained no seed like it and so appeared to support the stronger claim that one-row leverage is absent here altogether. Seed 3 has $\rho_{\mathrm{LOO}} = 1.16$ with exactly one flipping row, and that row is again a rare-class row on which the arms differ by 990.0, the same realised ESCALATE-to-IGNORE swing as on AI4I. So the two statistics separate on most of this domain but not all of it, which cautions against reading ρ_{LOO} as a restatement of the ceiling—and against reading a partial seed sweep as the domain.

Fidelity is a precondition, not a footnote. A statistic whose entire value is that it is exact cannot be computed on an inexact replay. The values here come from a re-run of the frozen protocol with per-row logging, and all three domains reproduce bitwise on $10/10$ seeds; an earlier cached-replay attempt reproduced the hard-label arm on $6/10$ and was discarded rather than reported.

5.2 How Much Evaluation Would Be Enough

ρ_{LOO} describes the sample obtained. The design question is how large a sample would have to be. A paired comparison is a mean of bounded differences $d_i \in [-D_{\max}, D_{\max}]$, so empirical Bernstein

Table 1: **Effect size against evaluation resolution, with the exact one-row sensitivity.** Δ is the paired mean RWL difference (confusion-matrix Bayes minus the best uncalibrated scalar-summary control) over 10 seeds. Point estimates are frozen-run values on both sides; the bootstrap supplies interval width only, and seed counts use per-seed 95% intervals (B=2000). "Rare rows" is the median rare-class count in the 500-row evaluation split. The *ceiling* column prices the most expensive single row the loss matrix admits—IGNORE on a FAILURE row, $L_{\max}/n_{\text{test}}$=2.0 RWL, whatever the prevalence—as a multiple of the measured effect. ρ_{LOO} (§5) answers what that ceiling can only bound: it exceeds 1 iff some single evaluation row can be deleted to reverse the empirical winner, and the next column counts the rows that do. The two disagree by an order of magnitude on NSL-KDD because the arms never realise the worst row the matrix allows—the point of computing the exact statistic rather than bounding it. The last column is descriptive and carries no error guarantee.

Domain	**vs.**	**Δ RWL**	**Seeds** (↓ better)	**Rare rows**	**ceiling** $/\lvert\Delta\rvert$	ρ_{LOO} (worst seed)	**flips** (rows)	**Reading**
NSL-KDD	wt. averaging	-1.511	10/10 ↓	64	1.3×	0.26	0	**decisive**
CICIDS2017	voting	-0.161	8/10 ↓, 1 ↑	27	12.4×	1.16	1	under-resolved
AI4I	voting	$+0.029$	8/10 ↓, 2 ↑	4	68.7×	1.29	2	*not adjudicable*

(Maurer & Pontil, 2009) gives, with probability $1-\delta$, the *one-sided* bound

$$\Delta \;\geq\; \widehat{\Delta} - h(n), \qquad h(n) = \sqrt{\frac{2V_n \log(2/\delta)}{n}} + \frac{14 D_{\max} \log(2/\delta)}{3(n-1)}, \tag{5}$$

with V_n the sample variance of d. One side is what a sign needs and is what Maurer & Pontil (2009) states at $\log(2/\delta)$; two-sided would cost $\log(4/\delta)$. The factor 14, rather than the 7 of the standard statement, is the rescaling: Maurer–Pontil is stated on $[0,1]$, a paired difference occupies $[-D_{\max}, D_{\max}]$ whose *width* is $2D_{\max}$, and pulling the bound back through $z = (d + D_{\max})/2D_{\max}$ multiplies both terms by $2D_{\max}$. The variance term absorbs that factor exactly, since $V_z = V_n/(2D_{\max})^2$; the range term does not. Both rescalings are machine-checked (App. J). The sign of Δ is resolvable *under this bound* only once $h(n)$ falls below $|\widehat{\Delta}|$; a sharper test could adjudicate sooner.

Proposition 1 (Observed-effect resolution requirement)**.** *Fix δ and hold the observed sample variance V_n and effect magnitude $|\widehat{\Delta}| > 0$ fixed. The half-width $h(n)$ of* (5) *is strictly decreasing in n, so there is a smallest n' at which it falls below $|\widehat{\Delta}|$, and it is unique and computable.*

The direction of any one-sided risk claim is fixed before evaluation, never chosen after seeing the sign. Two words of care. n' is a *retrospective* diagnostic, not a prospective guarantee: V_n and $|\widehat{\Delta}|$ are read off the comparison that was run, so n' answers "how many rows would this observed effect have needed", not "how many should we collect". And the monotonicity is load-bearing: we locate n' by binary search, sound only because the half-width decreases, and machine-check it in App. J.

The consequence is uncomfortable and worth stating plainly. On NSL-KDD, the domain we call decisive, a single seed's paired comparison would need 6.6–31k evaluation rows (6,579–31,317 across seeds, δ=0.05) before (5) could resolve its sign; on AI4I 21–213k, and on CICIDS2017 20–106k. The protocol uses 500. What carries the NSL-KDD conclusion is therefore *not* any single seed's interval; it is that all ten independent seeds agree in sign, which the exact sign-flip test evaluates at the seed level and which ρ_{LOO} shows is not the work of one row. These are different statements about different units—rows within a seed, and seeds—and reporting a single p-value invites conflating them.

On AI4I the requirement is finite but unreachable, and for a structural reason rather than a budgetary one: the domain's *entire* test split contains 19 rare examples, so no enlargement of n_{test} within this protocol approaches even the smallest per-seed requirement. Its aggregate $\Delta = +0.029$ is also not a small per-seed effect but a near-cancellation of much larger ones—per-seed Δ ranges over $[-0.55, +1.65]$—which is what *not adjudicable* means and what a single aggregate p-value hides.

What the audit refuted. Fitting a 4×4 confusion matrix from as few as nine rare calibration rows looks precarious, so we bootstrapped the calibration split (B=2000, whole rows, refit propagated

through the deployed decision path to a committed action). Calibration-sampling uncertainty is small—per-seed 95% intervals have mean width 0.11–0.23 RWL and the committed action changes on 0.07–1.56% of rows, with one CICIDS2017 seed spanning 1.989 as the exception—so it is rare-event *evaluation* granularity, not confusion-matrix fitting noise, that dominates. The reading column of Table 1 is computed by a fixed rule on the per-seed intervals alone— whether any seed's interval favours the control, and whether the aggregate sign agrees with the majority—enumerated in that caption. It is descriptive and carries no error guarantee, whereas the per-seed intervals it rests on were pre-specified.

6 A FINITE-SAMPLE REJECTION CERTIFICATE

Correctly specified Bayes fusion cannot be beaten on the same information, so the object of study is the *misspecified* rule actually running. What determines the cost of its failures is where its rare-class *actions* land.

Lemma 2 (Rare-class risk of the deployed rule)**.** *Let A be the action the deployed rule commits and $r_a^* = \Pr[A{=}a \mid Y{=}y^*]$ its action-destination distribution on the rare class. Then $\mathcal{R} \geq B^{y^*} := \pi_{y^*} \sum_a r_a^* L(a, y^*)$.*

We index by the *action* deliberately: the rule maps the whole belief to an action, so the action is not in general a function of the modal label, and a label-indexed version would not describe the system we run. Chaining against a comparator whose risk is bounded above gives the population condition.

Proposition 3 (Population rejection)**.** *Let g_c satisfy $\mathcal{R}(g_c) \leq U_c$. If $B^{y^*} > U_c$ then $\mathcal{R} > \mathcal{R}(g_c)$: the deployed rule is strictly worse and may be rejected. If $B^{y^*} \leq U_c$ the condition is uninformative and no conclusion follows in either direction.*

The gap this leaves. What ships estimates both sides on the calibration split through the deployed decision path,

$$\widehat{B}^{y^*} = \widehat{\pi}_{y^*} \cdot \frac{1}{n_{y^*}} \sum_{i\,:\,Y_i = y^*} L(A_i, y^*), \tag{6}$$

alongside a plug-in $\widehat{\mathcal{R}}(g_c)$, and thresholds the difference at zero. Substituting a point estimate for the upper bound U_c breaks Proposition 3's chain—a formal witness of the non-lifting is in the supplementary material—so the deployed object is a screen with no finite-sample guarantee. We now give the guarantee.

6.1 THE ESTIMAND, WRITTEN SO THE SAMPLE SIZE IS RIGHT

Equation (6) is a ratio of two estimates over n_{y^*} rare rows. Write instead

$$X_i = \mathbb{1}\{Y_i{=}y^*\}\, L(A_i, y^*) \in [0, L_{\max}], \qquad B^{y^*} = \mathbb{E}[X]. \tag{7}$$

This is not a cosmetic rewrite. It gives the *identical point estimate*—the indicator zeroes every non-rare row, so the sum is unchanged and the divisor becomes n—as a single mean of i.i.d. bounded variables over *all* n calibration rows. The concentration term is then $O(1/\sqrt{n})$ in the full calibration size rather than in the rare count: on NSL-KDD, 2000 rather than 251. What this buys is validity, not a like-for-like radius ratio: only the indicator form is a single i.i.d. mean rather than a product of two estimates.

6.2 THE CERTIFICATE

Let $C_j = L(A_j^{g_c}, Y_j) \in [0, L_{\max}]$ be the comparator's realised losses on the same split, with $\mathcal{R}(g_c) = \mathbb{E}[C]$.

Theorem 4 (One-sided finite-sample rejection)**.** *Let $(X_i)_{i \leq n}$ and $(C_j)_{j \leq m}$ be i.i.d. samples in $[0, L_{\max}]$ with means B^{y^*} and $\mathcal{R}(g_c)$. Fix $\delta \in (0, 1)$ and set*

$$\mathrm{LCB} = \bar{X} - L_{\max}\sqrt{\tfrac{\log(4/\delta)}{2n}}, \quad \mathrm{UCB} = \bar{C} + L_{\max}\sqrt{\tfrac{\log(4/\delta)}{2m}}.$$

Table 2: **What the deployed screen claims, and what a finite-sample certificate actually licenses** (δ=0.05, per seed, 10 seeds). $\widehat{B}$ and $\widehat{\mathcal{R}}(g_c)$ are the two point estimates the shipped screen compares; *plug-in* is their difference, which the screen thresholds at zero. The certified margins replace both by one-sided bounds at level $\delta/2$ and subtract, so a *positive* certified margin is a genuine $1-\delta$ statement that the deployed rule is worse (Thm. 4); a negative one abstains and asserts nothing. Hoeffding is the headline *because* it is the weaker bound, which is what makes "the bound was not chosen to fire" checkable. *Read the columns by their own label:* $\widehat{B}$ and $\widehat{\mathcal{R}}(g_c)$ are means over seeds, every *margin* is that margin's worst seed, and the three need not come from one seed—so the row is not a decomposition and $\widehat{B} - \widehat{\mathcal{R}}(g_c)$ is not the reported margin. On AI4I and CICIDS2017 the plug-in margin is itself negative, so neither the screen nor the certificate rejects.

				plug-in	**Hoeffding**		**emp. Bernstein**	
Domain	n_{cal}	$\widehat{B}^{y^*}$ (mean)	$\widehat{\mathcal{R}}(g_c)$ (mean)	**margin** (worst)	**margin** (worst)	**cert.**	**margin** (worst)	**cert.**
NSL-KDD	2000	92.92	13.97	+67.7	+1.5	**10/10**	+37.5	**10/10**
CICIDS2017	2000	0.63	4.40	−4.3	−70.5	0/10	−16.0	0/10
AI4I	1500	0.02	4.79	−4.8	−81.3	0/10	−18.6	0/10

Then

$$\Pr\left[\text{LCB} > \text{UCB} \;\textit{and}\; \mathcal{R} \le \mathcal{R}(g_c)\right] \;\le\; \delta,$$

so firing certifies $\mathcal{R} > \mathcal{R}(g_c)$ at level δ. What is controlled is the probability of declaring the deployed rule worse when it is not; nothing here bounds the probability of detecting a rule that is.

Proof sketch. One-sided Hoeffding gives $\Pr[\bar{X} - B^{y^*} \ge t] \le e^{-2nt^2/L_{\max}^2}$, which at $t = L_{\max}\sqrt{\log(4/\delta)/2n}$ is $\delta/4$; symmetrically for $\bar{C}$. On the complement of both failure events, which a union bound gives probability at least $1 - \delta/2 \ge 1 - \delta$, we have $B^{y^*} \ge \text{LCB} > \text{UCB} \ge \mathcal{R}(g_c)$, and Lemma 2 supplies $\mathcal{R} \ge B^{y^*}$. Applying Proposition 3 with $U_c = \text{UCB}$ completes it. □

Remark 5 (The constant is deliberately loose). A tight one-sided bound at level $\delta/2$ carries $\log(2/\delta)$, not $\log(4/\delta)$; we keep the larger constant, which is what the implementation uses, so each side fails with probability $\delta/4$ and the conclusion holds at $1 - \delta/2$. The slack is 9% in the half-width and it moves in the safe direction: every certified margin reported below is a lower bound on what the tight constant would give.

Substituting empirical Bernstein (5) for Hoeffding on both sides gives the same conclusion with a variance-adaptive half-width, and it is much tighter here because X_i is zero on 87–99% of rows so $V_n \ll L_{\max}^2$. **We report Hoeffding as the headline precisely because it is the weaker bound.** Reporting the conservative bound first is the only version of "we did not select a bound to make the certificate fire" that a reader can check, and the choice was fixed in advance; empirical Bernstein appears beside it and never in place of it.

What the hypotheses require. Theorem 4 needs (X_i) and (C_j) to be i.i.d. given the frozen trained pool, which would fail if either rule were fitted on the rows used to certify it. Neither is: both are fixed aggregators, every model in the pool is trained on the disjoint training split, and nothing is fitted on the certificate path. Conditioning on the pool is free, as a bound holding at $1 - \delta$ for each fixed pool holds at $1 - \delta$ on average over the training draw. Independence *between* the two samples is not required—a union bound does not need it—so evaluating both on the same rows is harmless. The one thing carried across rows is how far the controller has read into each oracle's latency stream, and that changes nothing: each query decision is measurable with respect to the current row and the draws revealed so far, so every unrevealed draw keeps its law and its independence. The per-row samples are therefore i.i.d. exactly, with no correction to δ (App. G).

6.3 What Fires and What Abstains

Table 2 reports all three levels of claim side by side. On NSL-KDD the certificate fires on $10/10$ seeds under Hoeffding, with a worst-seed certified margin of $+1.48$ RWL against a worst-seed

plug-in margin of $+67.68$: the gap between those two columns is the price of the guarantee, and the guarantee survives paying it. Empirical Bernstein certifies the same $10/10$ with a worst-seed margin of $+37.53$, showing where the headroom is—the comparator's half-width falls from 33.10 to 7.06 once variance is used.

Two design facts belong in the open. First, the certificate is evaluated with the comparator estimated on the *full* calibration split, whereas the shipped screen uses a 500-row subsample—a compute shortcut in the original code, not a statistical choice. At n_c=500 the certificate does *not* fire (LCB 64.12 against UCB 79.74); at n_c=2000 it does. This was discovered before the analysis plan was written down and is recorded as such in our pre-registration rather than presented as a prediction; it adds no data, changes no estimator, and the full-split comparator estimates (13.0–15.1) bracket the frozen subsample ones (11.9–15.8). Second, on AI4I and CICIDS2017 alike the certificate abstains on $10/10$ seeds under both bounds—but so does the plug-in screen, whose mean margins there are -4.78 and -3.78. Neither abstention is the bound failing to keep up with a real effect; there is no rejection to certify.

A non-firing certificate is still a result. The three-level structure (population condition, certified version, plug-in screen) is the contribution, not the fact that the middle level happens to fire on one domain. Where the certified margin is negative and the plug-in margin positive, the shipped screen is claiming more than the data supports, and reporting that is the point of computing it.

7 RELATED WORK

Budgeted inference varies *when* to query or *which* predictor acts (Viola & Jones, 2001; Bolukbasi et al., 2017; Chow, 1970; Geifman & El-Yaniv, 2017; Madras et al., 2018); all hold the fusion rule fixed. **Classifier fusion** varies *how*: the sum and vote rules differ in their sensitivity to estimation error because voting discretises posteriors (Kittler et al., 1998), majority-vote accuracy is bounded under dependence (Kuncheva et al., 2003), and confusion-matrix models of unequally competent experts go back to Dawid & Skene (1979); Nitzan & Paroush (1982); Berend & Kontorovich (2014). **Cost-sensitive learning** supplies the other half: under asymmetric loss the optimal decisions are cost-weighted threshold shifts and accuracy is the wrong criterion (Elkan, 2001; Domingos, 1999). **Concentration** for bounded means is classical; we use Hoeffding and the empirical Bernstein bound of Maurer & Pontil (2009), and our contribution is not the inequality but the estimand (7) it is applied to. Recent cascades and routers (Nie et al., 2024; Dekoninck et al., 2024; Kolawole et al., 2025; Rabanser et al., 2025), multi-expert deferral (Mao et al., 2023; Alves et al., 2024; DeSalvo et al., 2025), and language-model ensembles (Jiang et al., 2023; Wang et al., 2024) decide when to stop, which model acts, or how outputs combine. None holds the decision setting fixed while replacing scalar per-model reliability with held-out class-conditional fusion under asymmetric action loss.

8 DISCUSSION

Limitations. Theorem 4 never clears a rule and its abstention carries no information; its i.i.d. assumption comes from the calibration split and would need revisiting under shift; and it is stated for a *fixed* comparator, so selecting among several would need a union bound we do not take. Its tightness is limited by $L_{\max}$ as a range, so a domain whose rare-class risk sits far below $L_{\max}$ abstains at any practical n—AI4I is that case. Its support requirement is a *different* estimand from the confusion matrices': gated on rare-row count, a domain can supply usable matrices and still be excluded from the screen. Theorem 6 is per model and for hard-label channels; soft predictive distributions and dependent pools are open. Confusion-matrix fusion estimates $O(C^2)$ cells per model, an apparent low-support vulnerability our audit did not find binding, though we do not assume that transfers. Our domains are tabular.

Conclusion. Under asymmetric loss, what a fusion rule must know about each model is not how often it is right but where its errors go, and for a hard-label channel a scalar reliability represents that only under a symmetric error model—which deployed rules assume anyway. Restoring class-conditional structure removes most of that cross-domain reversal among scalar-summary rules, where the comparison resolves at all. On two of three domains it does not, and saying so beats a p-value that conflates the two.

REFERENCES

Jean V. Alves, Diogo Leitão, Sérgio Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Cost-sensitive learning to defer to multiple experts with workload constraints. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2403.06906.

Daniel Berend and Aryeh Kontorovich. Consistency of weighted majority votes. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *International Conference on Machine Learning*, pp. 527–536. PMLR, 2017.

Chi-Keung Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. *arXiv preprint arXiv:2410.10347*, 2024. doi: 10.48550/arXiv.2410.10347.

Giulia DeSalvo, Clara Mohri, Mehryar Mohri, and Yutao Zhong. Budgeted multiple-expert deferral. *arXiv preprint arXiv:2510.26706*, 2025.

Pedro Domingos. MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, 1999. doi: 10.1145/312129.312220.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pp. 973–978, 2001.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Y. S. Huang and C. Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):90–94, 1995. doi: 10.1109/34.368145.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2 edition, 2014. doi: 10.1007/978-1-4471-5361-0.

Steven Kolawole, Don Dennis, Ameet Talwalkar, and Virginia Smith. Agreement-based cascading for efficient inference. *Transactions on Machine Learning Research*, 2025.

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2 edition, 2014. doi: 10.1002/9781118914564.

Ludmila I. Kuncheva, Christopher J. Whitaker, Catherine A. Shipp, and Robert P. W. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6(1):22–31, 2003. doi: 10.1007/s10044-002-0173-7.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018. doi: 10.48550/arxiv.1711.06664.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0159.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online cascade learning for efficient inference over streams. In *International Conference on Machine Learning*. PMLR, 2024. doi: 10.48550/arxiv.2402.04513.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982. doi: 10.2307/2526438.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, and Federico Tombari. Gatekeeper: Improving model cascades through confidence tuning. In *Advances in Neural Information Processing Systems*, 2025.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 511–518. IEEE, 2001. doi: 10.1109/CVPR.2001.990517.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

SUPPLEMENTARY MATERIAL

A ROBUSTNESS

The NSL-KDD result survives six perturbations, enumerated in the supplement: a heterogeneous pool, three alternative orchestrators, measured wall-clock cost, an n-sweep, a pool-size sweep and hard-deadline stress. One is worth naming: walking the loss matrix moves the CICIDS2017 winner from voting to Bayesian at $L_{\max} \geq 8000$ while rare-class accuracy is unchanged at 0.998—Lemma 8's radius shrinking, measured rather than derived.

B THE LOSS MATRIX

Table 3 is the matrix used throughout. Its two load-bearing entries are quoted in the main text where they are first needed—1000 for missing a failure against 10 for over-escalating on a normal input, a 100:1 ratio—so the table is here rather than in §2.

Table 3: **Asymmetric loss matrix $L(a, y)$ used throughout (lower is better).** Missing a failure (`ignore`/`failure`, bold) costs 1000 against 10 for over-escalating on a normal input (`escalate`/`normal`), a 100:1 asymmetry; this is what drives the rare-class regime in which the scalar-summary reversal occurs, and it is also why a single missed rare-class row can dominate an evaluation.

Action ↓ / State →	**normal**	**degraded**	**critical**	**failure**
ignore	0	10	100	**1000**
monitor	5	0.5	10	100
escalate	10	5	1	10
emergency	15	10	5	1.5

C WHY BITWISE FIDELITY NEEDS A QUIESCENT MACHINE

XGBoost's histogram construction is parallel, and the hard-label arms consume a discrete product, so one flipped training split can flip one label and move them while leaving the continuous arms untouched. Fixing the thread count is necessary—an unpinned re-run reproduces the soft arm exactly and the hard arm only to 0.008 RWL—but it is not sufficient. Across our executions of one seed set at *identical* thread count and affinity, every seed that completed in under 500 s reproduced bitwise, while the two that ran long under load from unrelated jobs on the same machine (726 s and 1972 s) did not, drifting by 0.019 and 0.051 RWL. The per-round deadline in our controller is charged against simulated oracle latency, not wall-clock, so this is not the controller reacting to load. What the evidence identifies is load-associated nondeterminism in *training*, consistent with parallel floating-point reduction order in histogram construction; we did not instrument the reduction itself, so we do not claim that mechanism uniquely. Operationally the consequence is the same either way: in our environment loaded runs failed bitwise reproduction even at fixed affinity, so we require a quiescent machine for an accepted re-run and reject any seed failing the frozen-output check. Wall-clock is only an early-stopping heuristic; the acceptance criterion is bitwise fidelity, never speed. We state this because a re-run that silently drifts by 0.05 RWL would be indistinguishable, to a reader, from one that agreed.

D THE CROSS-DOMAIN REVERSAL AMONG SCALAR-SUMMARY RULES

Table 4 is the symptom §1 quotes: among rules that see only hard labels or a scalar accuracy, which rule is best flips across domains. It is *prior* work on this system, not a contribution of this paper, which is why it sits here; the two swings the introduction cites are its NSL-KDD and AI4I columns.

Where Arm B sits in the fusion taxonomy. Arm B is not a new rule and we claim nothing for it. Estimating each model's confusion matrix on labelled data and combining as $\pi(y)\prod_k P(\hat{y}_k \mid y)$ is the *naive Bayes combiner*, also called the independence model, of the classifier-fusion literature (Kuncheva, 2014); the per-expert confusion matrix itself goes back to Dawid & Skene (1979). The factorisation is what makes it naive: the unfactorised alternative estimates $P(y \mid \hat{y}_1, \ldots, \hat{y}_K)$ on the joint output vector directly, which is the Behaviour Knowledge Space combiner (Huang & Suen, 1995), and needs support exponential in K that our five-model pools on a rare class do not have. What this paper contributes is not the rule but the matched control of §4 that isolates what the rule's class-conditional structure is worth, the resolution analysis of §5, and the certificate of §6.

Table 4: **What the fusion rule must know. Among rules that see only hard labels or a scalar accuracy, the best one reverses across domains; a rule that uses each model's full class-conditional error structure is decisively better where the rare class is adequately sampled (Table 1).** Mean RWL over n=10 paired seeds sharing the orchestration, pool and data order, so the rows differ only in how model outputs become a belief. Lower is better. **Bold** = best in column; underlined = not separable from the best by an exact two-sided sign-flip test at α=0.05 (p-value floor $2/2^{10}$=0.002). Temperature is fitted on calibration NLL alone and the aggregation rule is left unchanged, so the "+ temperature" row is a *sharpness-only* intervention: sharpening alone does not close the gap. It does *not* separate class-conditional structure from the hard-versus-soft change of representation the confusion row also makes (§4). Confusion matrices and the class prior are estimated on the held-out calibration split, with Laplace smoothing fixed in advance; protocol and reading were specified before the run (supplementary, S1).

Fusion rule	Uses	NSL-KDD	AI4I	CICIDS2017
Majority voting	hard labels	102.05	**0.88**	1.28
Reliability-Bayes	scalar accuracy	13.80	4.86	4.53
Weighted averaging	scalar accuracy	7.45	8.70	7.69
+ temperature	+ sharpness	7.01	4.90	3.01
Confusion-matrix Bayes	**full** $\widehat{P}(\hat{y} \mid Y)$	**5.93**	0.91	**1.12**

E SCALAR RELIABILITY AS A CHANNEL MODEL

The statements §3 summarises, with their discussion. Model one source as a channel $Q_{yc} = \Pr[Z{=}c \mid Y{=}y]$ over hard labels, modelled by a single accuracy α through the symmetric channel S_α of (1).

Theorem 6 (Scalar sufficiency, per model)**.** *Fix $C \geq 2$, a full-support prior, and $\alpha \in (0,1)$ with $\alpha \neq 1/C$. If the posterior induced by S_α equals the posterior induced by Q for* every *observed label, then $Q = S_\alpha$. The converse is immediate.*

The proof reduces to the invertibility of $S_\alpha = (\alpha - \beta)I + \beta J$, whose non-trivial eigenvalue $(\alpha C - 1)/(C-1)$ vanishes precisely at $\alpha{=}1/C$—the same non-degeneracy under which the label-noise literature inverts a noise-transition matrix to correct a loss (Natarajan et al., 2013; Patrini et al., 2017; van Rooyen & Williamson, 2018). Two qualifications are load-bearing. At $\alpha{=}1/C$ the channel is uninformative and the characterisation genuinely fails. And necessity does *not* lift to a product of K models, for which we give an explicit counterexample in the supplementary material; the theorem is per model, and the experiments fuse five.

Corollary 7 (Equal reliability, different action)**.** *For every $C \geq 2$ there exist channels $Q \neq Q'$ with identical overall accuracy, an observed label, and a loss under which their Bayes-optimal actions differ. No rule reading only α can be Bayes-optimal for both.*

A witness: with $\pi = (0.99, 0.01)$ and a binary model accurate on NORMAL but at chance on the rare class, the true and symmetric-model posteriors on $Z{=}$NORMAL are 0.00558 and 0.00117, straddling the IGNORE/MONITOR boundary $5/905{=}0.005525$. Both channels have accuracy exactly 0.896, so no scalar rule separates them, yet they commit different actions at an excess conditional risk of 0.050 per round.

Lemma 8 (Loss-boundary stability)**.** *Let p be the true posterior with unique Bayes action a^* and margins $\gamma_a = (L_a - L_{a^*})^\top p > 0$. For any $\tilde{p}$ on the simplex, $\|p - \tilde{p}\|_1 < \min_{a \neq a^*} 2\gamma_a / \operatorname{span}(L_a - L_{a^*})$ implies $\arg\min_a L_a^\top \tilde{p} = a^*$.*

Misspecification is free until it is not: a perturbation costs the committed action nothing until it crosses a loss-induced margin, then costs the full action gap—and the radius shrinks as the loss sharpens, which is why the asymmetry is not incidental. The algebraic statements here are machine-checked in Lean 4, which verifies derivational correctness but not the applicability of the assumptions.

F PROOFS

Every statement below is also machine-checked; App. J maps each to its Lean 4 name. Lean verifies derivational correctness, not the applicability of a hypothesis.

Proof of Theorem 6, scalar sufficiency per model. Write $P_M(y \mid c) \propto \pi(y) M(y, c)$. Equality of the posteriors induced by Q and S_α at every observed label c means, after cancelling $\pi(y) > 0$, $Q(y, c) = \lambda_c S_\alpha(y, c)$ for all y, with $\lambda_c = \sum_{y'} \pi(y') Q(y', c) \big/ \sum_{y'} \pi(y') S_\alpha(y', c) > 0$ depending on c alone. Both channels are row-stochastic, so summing over c gives $\sum_c \lambda_c S_\alpha(y, c) = 1$ for every y; since also $S_\alpha \mathbf{1} = \mathbf{1}$, the vector $\lambda - \mathbf{1}$ lies in $\ker S_\alpha$. Now $S_\alpha = (\alpha - \beta) I + \beta J$ with $\beta = (1 - \alpha)/(C - 1)$ and J all-ones, whose eigenvalues are 1 on $\mathbf{1}$ and $\alpha - \beta = (\alpha C - 1)/(C - 1)$ on its complement. As $\alpha \neq 1/C$ this is non-zero, S_α is invertible, $\lambda \equiv 1$, and $Q = S_\alpha$. The converse is immediate. At $\alpha = 1/C$ the kernel is $(C - 1)$-dimensional and the characterisation genuinely fails. □

Proof of Corollary 7, equal reliability different action. A witness suffices, and we give it in exact rationals. Take $C = 2$ with $\pi = (\frac{99}{100}, \frac{1}{100})$ and a channel accurate on NORMAL but at chance on the rare class, $Q(\text{N}, \text{N}) = \frac{9}{10}$, $Q(\text{F}, \text{N}) = \frac{1}{2}$. Its overall accuracy is $\alpha = \frac{112}{125} = 0.896$, and the symmetric channel S_α carries the same accuracy by construction. On $Z =$ NORMAL the two posteriors on the rare class are $\frac{5}{896}$ and $\frac{13}{11101}$. Under the loss matrix of Table 3 the IGNORE/MONITOR indifference point is $\frac{5}{905}$, which lies strictly between them. So the two channels commit different actions while sharing an accuracy, and the excess conditional risk of the scalar model's action is $\frac{45}{896} \approx 0.050$ per round. Since a rule reading only α receives identical input in both cases, it cannot be Bayes-optimal for both. □

Proof of Lemma 8, loss-boundary stability. Let $\delta = \tilde{p} - p$, so $\sum_i \delta_i = 0$ because both lie on the simplex. Fix $a \neq a^*$ and put $d = L_a - L_{a^*}$, so $d^\top p = \gamma_a > 0$ and $d^\top \tilde{p} = \gamma_a + d^\top \delta$. Because δ is zero-sum, $d^\top \delta = (d - k\mathbf{1})^\top \delta$ for every k; choosing $k = \frac{1}{2}(\max_i d_i + \min_i d_i)$ gives $\|d - k\mathbf{1}\|_\infty = \frac{1}{2} \operatorname{span}(d)$. Hölder then yields $|d^\top \delta| \leq \frac{1}{2} \operatorname{span}(d) \|\delta\|_1 < \gamma_a$ under the stated radius, so $d^\top \tilde{p} > 0$. As this holds for every $a \neq a^*$, the perturbed Bayes action is still a^*. □

Proof of Proposition 1, observed-effect resolution. On $n > 1$ both terms of (5) are strictly decreasing: $n \mapsto \sqrt{2V_n \log(2/\delta)/n}$ because $n \mapsto 1/n$ is, and $n \mapsto 14 D_{\max} \log(2/\delta)/(3(n-1))$ likewise. Their sum is therefore strictly decreasing and tends to 0, so $\{n : h(n) < |\widehat{\Delta}|\}$ is an up-set, non-empty, with a unique least element n'. Strict monotonicity is exactly the predicate a binary search needs, which is how n' is computed. □

Proof of Lemma 2, rare-class risk of the deployed rule. Conditioning on the true state and discarding the non-rare terms, which are non-negative because $L \geq 0$,

$$\mathcal{R} = \mathbb{E}[L(A, Y)] = \sum_y \pi_y \, \mathbb{E}[L(A, y) \mid Y{=}y] \;\geq\; \pi_{y^*} \mathbb{E}[L(A, y^*) \mid Y{=}y^*],$$

$$\pi_{y^*} \mathbb{E}[L(A, y^*) \mid Y{=}y^*] = \pi_{y^*} \sum_a r_a^* L(a, y^*) = B^{y^*}.$$

The index is the *action* because $\mathbb{E}[L(A, y^*) \mid Y{=}y^*]$ is an expectation over the action-destination distribution r^*; a label-indexed version would describe a different rule. □

Proof of Proposition 3, population rejection. Chain Lemma 2 with the hypotheses: $\mathcal{R} \geq B^{y^*} > U_c \geq \mathcal{R}(g_c)$. If instead $B^{y^*} \leq U_c$ the chain is broken at its middle link and no ordering of $\mathcal{R}$ and $\mathcal{R}(g_c)$ follows; both are realisable, which is why the screen abstains rather than concluding a tie. □

Proof of Theorem 4, one-sided finite-sample rejection. One-sided Hoeffding for a mean of i.i.d. variables in $[0, L_{\max}]$ (e.g. Vershynin, 2018, Thm. 2.2.6) gives $\Pr[\bar{X} - B^{y^*} \geq t] \leq \exp(-2nt^2/L_{\max}^2)$, which at $t = L_{\max}\sqrt{\log(4/\delta)/2n}$ equals $\delta/4$; symmetrically $\Pr[\bar{C} - \mathcal{R}(g_c) \leq -t'] \leq \delta/4$ at the matching t'. A union bound puts both complements simultaneously in force with probability at least $1 - \delta/2 \geq 1 - \delta$. On that event $B^{y^*} \geq \mathrm{LCB}$ and $\mathcal{R}(g_c) \leq \mathrm{UCB}$, so the firing condition $\mathrm{LCB} > \mathrm{UCB}$ gives $B^{y^*} > \mathcal{R}(g_c)$, and Proposition 3 with $U_c = \mathrm{UCB}$, together with Lemma 2, gives $\mathcal{R} > \mathcal{R}(g_c)$. □

G THE I.I.D. HYPOTHESIS ON THE CERTIFICATE PATH

Theorem 4 is only as good as its hypotheses, and the way such a certificate usually fails is that the rule being certified was fitted on the certifying rows. We audited the shipped path for this and report the audit rather than an assurance.

Nothing is fitted on the certificate sample. The deployed rule is plain voting and the comparator is Reliability-Bayes. Both are fixed aggregators instantiated from a method name, and this is worth stating precisely because the comparator's name suggests otherwise: on this path its per-model reliabilities are held at the uniform construction-time value $1/2$ and are never updated, so "Reliability-Bayes" here is a likelihood product at constant weights, not a rule carrying fitted reliabilities. The reliability, temperature and confusion-matrix fitting that the fusion arms of §4 perform is not invoked anywhere on this path. Every model in the pool is trained on the training split, which is disjoint from the split the certificate uses. Theorem 4 is a statement about a *fixed* comparator, and this is the comparator it is about; fitting reliabilities would define a different comparison with its own certificate, not a sharper reading of this one. We make no claim about which of the two carries lower risk. Note also that (X_i) and (C_j) need not be independent of *each other*: a union bound over the two failure events does not require it, so estimating both on the same rows is legitimate.

Conditioning on the frozen pool. Every certificate row is scored by one trained pool θ, so the X_i are not *unconditionally* independent: they share whatever the training draw put into θ. Both the statement we need and the statement we make are conditional on θ. At fixed θ, $X_i = f_\theta(\mathrm{row}_i, L_i)$ is a fixed measurable function of an i.i.d. pair, so Theorem 4 applies verbatim and holds at level $1 - \delta$; since the same δ works for every θ, integrating over the training randomness returns the same $1 - \delta$. The conditional reading is also the deployment-relevant one, because θ is estimated on a split disjoint from the certificate split. What conditioning does *not* buy is a claim about a fresh training draw: Table 2 reports a per-seed certificate at a fixed pool, not a bound that survives retraining.

What is carried across rows. What would break the hypothesis is a rule whose behaviour on row i depends on rows $j < i$. We enumerated the controller's mutable state. The budget and the deadline are reset at every call, so neither persists. The per-model reliabilities are constants. Predictions are deterministic functions of the features. Two counters are incremented but are read only by a reporting method, never by the decision logic. That leaves exactly one thing: the position in each oracle's latency stream.

Why the shared stream does not break independence. Each oracle k carries its own i.i.d. latency stream $\ell_1^{(k)}, \ell_2^{(k)}, \ldots$; the K streams are mutually independent and all are seeded independently of the evaluation rows. On row i the controller reads a block L_i from these streams, and how far it reads into each is decided adaptively: which oracle to query next may depend on the current row and on latencies already revealed from *other* oracles. We therefore do not claim that the index reached in any single stream is a stopping time for that stream's own filtration—adaptive cross-stream querying need not make that true—and argue against the joint revealed history instead. Let $\mathcal{F}_t$ be the σ-algebra generated by the rows seen so far and by every latency value revealed so far, across all streams. Conditional on the frozen pool, the current row is independent of all unrevealed draws, and each query decision is $\mathcal{F}_t$-measurable by construction. Hence at every reveal time the next unused draw of each stream retains its original law and is independent of $\mathcal{F}_t$: adaptively choosing *which*

Table 5: **The matched control: scalar reliability against class-conditional structure, with the representation held fixed.** Both arms fuse hard labels as $\pi(y)\prod_k P(\hat{y}_k \mid y)$ through the same code path, rows, calibration split and Laplace smoothing. The only difference is the channel: the symmetric $S_{\hat{\alpha}_k}$ built from oracle k's own accuracy, against its fitted confusion matrix Q_k. Δ is the paired mean RWL difference $S - Q$, so positive favours class-conditional structure; p is the exact two-sided sign-flip test on paired per-seed differences, floor $2/2^{10}$=0.002 at n=10. *Fidelity* counts seeds on which the re-run Q arm reproduced the frozen protocol *bitwise*; seeds that did not are excluded from every other column rather than averaged in, and a count below 10 is bold so the exclusion cannot be missed. The reading of both outcomes was fixed before the run.

Domain	$S_{\hat{\alpha}}$ RWL	Q RWL	Δ $S-Q$	**Seeds** Q better	**Sig.**	**Fid.** exact
NSL-KDD	112.295	5.935	+106.360	10/10	p=0.002	10/10
CICIDS2017	1.319	1.118	+0.200	4/10	p=0.188	10/10
AI4I	2.061	0.906	+1.154	9/10	p=0.010	10/10

independent stream to read next cannot bias what that stream returns. Equivalently, the joint latency process is Markov and the row boundary is a stopping time for it, so the strong Markov property (Klenke, 2014, Thm. 17.14) applies to the block as a whole. So L_i is fresh and identically distributed however row i was reached, the pairs (row_i, L_i) are i.i.d., and $X_i = \mathbb{1}\{Y_i{=}y^*\}L(A_i, y^*)$—a fixed measurable function of that pair at fixed pool—is i.i.d. as Theorem 4 requires. **No correction to δ is needed**, and the guarantee reported in Table 2 is the stated $1 - \delta$.

A stronger statement we can also make. Independent of the argument above, the latency block almost never matters at all. A draw can alter a committed action only by tripping a deadline gate, which refuses an oracle when its nominal cost exceeds a fraction of the deadline remaining; with the deployed pool and a 500 ms deadline that needs the realised latency of the first four queries to exceed 262.5 ms against a mean of 105.9 ms. Numerically integrating the latency law gives 2.1×10^{-12} per row, or 1.2×10^{-7} over the 60,000 rows the certificate uses. So A_i is not merely a function of (row_i, L_i) but, outside an event of that probability, of row_i alone. We report the bound because it prices how much the latency model is doing, which is nothing; the certificate does not rest on it.

We also record a false lead, because it is the kind a reader may worry about. The controller does mutate a value-of-information estimator across rows. That state is inert: both branches of the escalation test return the same next oracle and differ only in the reason label they attach, so the mutated statistic cannot reach the action.

H DEPLOYMENT, ORDERED BY THE EVIDENCE

(1) With labelled calibration data of adequate rare-class support, fuse with confusion structure. **(2)** Audit an inherited rule with Theorem 4 rather than with the plug-in screen, and treat a non-trigger as *no information* rather than clearance. **(3)** Before reading a non-significant comparison as a tie, compute ρ_{LOO} and Proposition 1; a non-significant test does not distinguish absence of an effect from insufficient resolution.

I THE MATCHED CONTROL, PER SEED

Table 6 applies the exact one-row sensitivity of (4) to the matched contrast itself, per domain and seed. It is analysis of data the E1 run already stored—the per-row paired difference for all 500 evaluation rows of every seed—so nothing here is a further experiment and no parameter is chosen. The statistic is imported from the same module that produces Table 1, so it is identical by construction rather than by description.

Two columns appear because neither suffices alone. A contrast decided by exactly one row has $\rho_{\mathrm{LOO}} = 1$ *exactly*, and the strict $\rho_{\mathrm{LOO}} > 1$ test reads that as stable; the differing-row count is what exposes it. Five of the ten CICIDS2017 seeds are that case and four more have no differing row at all, which is why that domain's aggregate is reported but never read as a replication or a refutation.

Table 6: **The matched control read through the resolution framework of §5, per seed.** Δ is the paired mean RWL difference $S_{\hat{\alpha}} - Q$ over that seed's 500 evaluation rows; ρ_{LOO} is the exact one-row sensitivity (4) applied to *this* contrast, so $\rho_{\mathrm{LOO}} > 1$ means one row can reverse that seed. *Differ* counts rows on which the two arms commit different actions at all. The two columns are both needed: a contrast decided by exactly one row has $\rho_{\mathrm{LOO}} = 1$ *exactly*, which the strict test reads as stable, so the differing-row count is what makes that case visible. On NSL-KDD the effect is spread over roughly 60% of rows and sits about a factor of 50 from a one-row flip on every seed. On AI4I two seeds are one row from flipping. On CICIDS2017 the arms differ on at most two rows and agree *exactly* on four seeds, so that domain has almost no contrast for the evaluation to resolve—which is why the protocol reads it here and never as a third vote.

Domain	**Seed**	Δ $S-Q$	ρ_{LOO} > 1 flips	**Differ** of 500	**Reading**
NSL-KDD	0	+124.4930	0.0160	313	—
	1	+104.4500	0.0191	301	—
	2	+113.4700	0.0176	289	—
	3	+94.6670	0.0211	284	—
	4	+105.8350	0.0189	307	—
	5	+93.2530	0.0214	301	—
	6	+97.0810	0.0206	311	—
	7	+116.6750	0.0171	316	—
	8	+109.3780	0.0183	302	—
	9	+104.3020	0.0191	302	—
AI4I	0	+5.1150	0.3904	24	—
	1	+0.2580	0.7674	12	—
	2	+0.6280	0.2866	16	—
	3	−0.0600	0.1667	6	—
	4	+0.0240	7.5000	27	**one row reverses it**
	5	+2.0510	0.9737	23	—
	6	+0.0180	10.0000	18	**one row reverses it**
	7	+0.8360	0.2368	18	—
	8	+0.4500	0.4000	12	—
	9	+2.2230	0.8983	20	—
CICIDS2017	0	+0.0180	1.0000	1	one-row effect
	1	−0.0100	1.0000	1	one-row effect
	2	+0.0000	*n/a*	0	arms agree exactly
	3	+0.0180	1.0000	1	one-row effect
	4	+0.0000	*n/a*	0	arms agree exactly
	5	+0.0000	*n/a*	0	arms agree exactly
	6	+0.1970	1.0000	1	one-row effect
	7	+1.7900	1.0056	2	**one row reverses it**
	8	−0.0100	1.0000	1	one-row effect
	9	+0.0000	*n/a*	0	arms agree exactly

USE OF LARGE LANGUAGE MODELS

This is the mandatory AI-use disclosure. We answer ICLR's required and recommended task lists first, then elaborate, because the lists alone do not convey how the assistance was checked.

Required-disclosure tasks for which we used generative AI. Helping develop theoretical models and conceptual frameworks; formulating mathematical claims; providing critical ingredients for proving them; assisting in the writing of proofs; proposing and refining hypotheses; designing and giving feedback on research methodology and experiments; implementing methods; cleaning and reformatting datasets, including the loaders and the mapping to the four-class severity scheme; and interpreting results.

Required-disclosure tasks that do not apply here. We generated no synthetic datasets: the three corpora are public, and the loss matrix is produced by a deterministic function in the released code

rather than by a model. This work involved no translation and no qualitative or thematic data analysis.

Recommended-disclosure tasks for which we used generative AI. Creating and modifying figures; creating and editing software code; drafting parts of this paper; editing it for readability; suggesting its structure; proposing its title; searching for and summarising related literature; identifying relevant work and formatting references; and brainstorming.

Writing and editing. LLM assistance was used extensively to draft and revise prose, restructure sections, and copy-edit. All scientific claims and their framing were checked and are owned by the authors.

Code. LLM assistance was used to implement and refactor parts of the experimental harness — dataset loaders, baseline wrappers, the table generators under `scripts/`, and the analysis scripts. All such code was executed and its output checked against the results files it reads. No number in this paper is typed by hand: every table is emitted by a generator, and an automated provenance gate re-derives each reported figure from the run that produced it and fails on drift.

Analysis. LLM assistance was used to summarise and cross-check experimental output. Every quantitative claim is traced to a specific results file by that gate rather than to any model's summary.

Theory. The algebraic content of Thm. 6, Cor. 7, Lem. 8, Lem. 2, Prop. 3, Prop. 1 and Thm. 4 is machine-checked in Lean 4 against Mathlib, with arithmetic side conditions discharged by z3; App. J gives the correspondence. Machine checking establishes derivational correctness, not that a hypothesis applies to the deployment. LLM assistance was used in developing and refining these statements and their proofs, and in checking algebra and prose.

Verification, and who is responsible. We have reviewed all AI-assisted work. Every table and figure is emitted by a generator from a stored result file, and an automated provenance gate re-derives each reported quantity from the run that produced it and fails on drift, so no number here is typed by hand. The theorem statements and proofs are machine-checked in Lean 4 against Mathlib with arithmetic side conditions discharged by z3, and reference metadata was verified against publisher records rather than taken from model output. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

Errors found and kept visible. Checking of this kind found errors that are corrected here rather than silently repaired, and are documented where they occur: an empirical-Bernstein range constant that was wrong by a factor of two (App. F and the rescaling lemmas of App. J), a table whose columns mixed per-seed means with worst-seed margins, and a resolution reading for CICIDS2017 that a partial three-seed sweep supported and the completed ten-seed sweep overturned (§5).

Responsibility. The authors have reviewed all AI-assisted work and take full responsibility for the contents of this submission.

J FORMAL VERIFICATION

Each statement is machine-checked in Lean 4 against Mathlib, with the arithmetic side conditions discharged by z3. The table gives the correspondence. Nothing here certifies that the modelling assumptions hold of the deployment; it certifies that the derivations are correct given them.

Statement	Lean name	File
Thm. 6	`scalar_sufficiency`	`ScalarSufficiency.lean`
$\alpha{=}1/C$ failure	`chance_level_exception`	`ScalarSufficiency.lean`
Cor. 7	`acc_eq`, `separates`, `excess`	`CorollaryInstance.lean`
Lem. 8	`margin_survives`	`LossBoundary.lean`
Lem. 2	`rare_class_lower_bound`	`LossBoundary.lean`
Prop. 3	`veto_chain`	`LossBoundary.lean`
non-lifting witness	`veto_chain_needs_upper_bound`	`LossBoundary.lean`
product counterexample	`joint_distribution_differs`	`NonLifting.lean`
Prop. 1	`eb_halfwidth_strict_anti`	`ResolutionCertificate.lean`
ρ_{LOO} characterisation	`loo_reversal_iff`	`ResolutionCertificate.lean`
Eq. (5) rescaling	`variance_rescale`, `range_term_rescale`	`ResolutionCertificate.lean`
Thm. 4 chain	`cert_chain`	`ResolutionCertificate.lean`
Hoeffding inversion	`hoeffding_inversion`	`ResolutionCertificate.lean`
Rem. 5	`loose_constant_safe`	`ResolutionCertificate.lean`

Two of these were written because the audit that produced this appendix found a gap rather than confirmed one. `eb_halfwidth_strict_anti` discharges a monotonicity our implementation had asserted in a code comment and relied on for a binary search. `variance_rescale` and `range_term_rescale` record the $[-D_{\max}, D_{\max}]$ rescaling of (5); an earlier version of this paper carried the range constant valid for a variable on $[0, D_{\max}]$, which understated the required sample sizes reported in §5. The values there are those of the corrected constant.

SUPPLEMENTARY MATERIAL (SEPARATE ARCHIVE)

The archive accompanying this submission is indexed **S1**–**S7**, numbered so as not to collide with the appendix sections above. Every table and figure in this paper is emitted by a generator in `scripts/` from a file in `results/`; the archive ships both, and `PROVENANCE.txt` is the output of the gate that re-derives each reported quantity from the run that produced it.

- **S1** Falsification pre-specification, its amendments, and the sanity endpoint that failed and whose tolerance was not widened.
- **S2** The finite-sample audit: protocol, fidelity record—including the cached-replay offset and why point estimates are not taken from it—per-seed 95% intervals, and action-change rates.
- **S3** Robustness: three alternative orchestrators, an n-sweep, hard-deadline stress, and the loss-matrix walk of §8.
- **S4** The matched control of §4, per seed and per row.
- **S5** The exact one-row sensitivity and the retrospective sample-size calculation of §5.
- **S6** The certificate of §6: per-seed margins under both concentration bounds, and the abstentions.
- **S7** Secondary study, marked non-headline: the language-model pool.