WHEN ALL LLM JUDGES AGREE: SILENT CONSENSUS MAKES DEPLOYED RISK UNMEASURABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judge panels now *gate* model outputs: unanimous approvals ship automatically, and humans see only disagreements. This design manufactures its own evidence of reliability—when all judges share a blind spot, the bad output ships and no label is ever collected, a regime we call *silent consensus*. We first prove the error rate of never-audited released outputs is unidentifiable: the same logs arise whether it is 0%, 100%, or anything in between. We then measure how often whole panels fail together, with frozen open-weight judges on three tasks with independent labels (policy compliance, logical reasoning, medical advice). Joint failures are common where judgment is subjective: a panel of three model sizes from one family unanimously approved 42.9% of medical responses violating physician-written criteria, and three samples of one model erred almost identically, jointly approving 39% of corrupted reasoning claims—a three-judge panel in name only. Finally, we compare auditing strategies in simulated deployments where the case mix shifts partway through (e.g., unseen expert criteria). Two natural strategies fail. Auditing only where judges disagree never checks a shared blind spot, by definition. Auditing the least-confident cases barely helps: judges are confident precisely where they are blind, and 68.5% of unsafe outputs released after the shift could never be audited. Auditing a small random fraction of released outputs, with logged probabilities, makes the error rate estimable, with a confidence bound valid under continuous monitoring. Randomized audits do not make a system safe, but without them deployed risk cannot be measured at all.

1 INTRODUCTION

LLM-as-a-judge has moved from benchmarks to pipelines (Zheng et al., 2023; Verga et al., 2024). A now-common deployment pattern places a panel of LLM judges between a generator and a consequential action: each candidate output—a code change, an analysis, a customer-facing response—is reviewed by several judges, unanimous approvals are released automatically, and only disagreements are escalated to scarce human reviewers. The appeal is obvious: generation parallelizes, human review does not, and escalating only disagreements concentrates human attention where the panel itself signals uncertainty.

This paper is about what that design does to the organization's ability to *measure* its own risk. When judges disagree, a human looks at the case and produces a label. When judges unanimously approve, the output ships and no one ever looks. The set of cases that receive ground-truth labels is therefore chosen by the very policy being evaluated—a *selective labels* problem (Lakkaraju et al., 2017)—and the unlabeled region is exactly the unanimous-approval region. If all judges share a blind spot, the bad output ships, the logs stay clean, and the pipeline's dashboards report the resulting silence as reliability. We call this regime **silent consensus**: unanimous approval that no independent label ever touches (Figure 1). Adding more judges from the same lineage makes the problem *look* better—disagreement drops, escalations drop—while leaving jointly missed errors untouched: LLM errors are strongly correlated within and even across model families (Kim et al., 2025).

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.

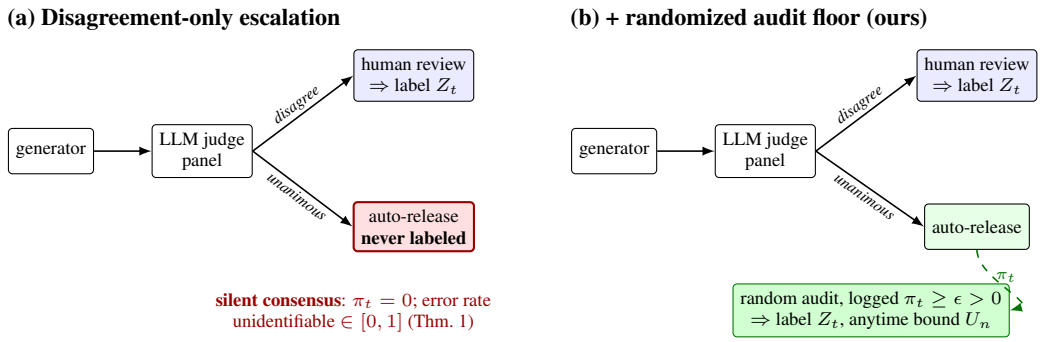


Figure 1: The problem and the fix at a glance. **(a)** A panel that escalates only disagreements labels only cases it already doubts; unanimous approvals—including every case where all judges share a blind spot—ship with zero audit probability, so their error rate is unidentifiable from the log (Theorem 1). **(b)** A small logged random audit floor $\pi_t \geq \epsilon > 0$ restores an independent label channel, making the released population's cumulative error estimable with an anytime-valid bound U_n; targeted auditing may be layered on top but can never zero out a stratum.

Two questions follow. First, is this a bias that downstream statistics can correct, or something worse? Second, how large is silent consensus in practice, and what does it take, operationally, to make deployed risk measurable again?

Our answer to the first question is an impossibility result. If a subpopulation of released outputs—a *stratum*, in the language of stratified sampling—has zero probability of ever being audited, the organization cannot tell which of two worlds it is in: one where every such output is safe, one where every one is unsafe. The logs are identical in both, so the unsafe rate is not nonparametrically point identified and the identified set is $[0, 1]$ (Theorem 1). This transplants classical results (Rubin, 1976; Rosenbaum & Rubin, 1983; Horvitz & Thompson, 1952; Manski, 1990)—positivity and worst-case identification bounds—to a missingness mechanism that is itself an LLM judge panel, with a blunt corollary: disagreement-only escalation can never certify the population it releases, however many disagreement labels it collects.

Our answer to the second question is empirical and systems-level. We measure *joint false-accept rates*—how often every judge on a panel approves an output that is actually bad—on frozen panels of pinned open-weight models across three tracks with independent ground truth (executable policy checks, executable logic checks, and physician-majority medical labels), varying panel composition along controlled axes: repeated samples of one model, paraphrased prompts, same-family checkpoints, and cross-family models. We then implement and stress-test a randomized audit protocol that enforces a logged, positive audit probability in every stratum of released outputs, estimates cumulative shipped errors by inverse propensity weighting, and maintains an anytime-valid upper confidence bound (Howard et al., 2021; Ramdas et al., 2023) that remains correct under adaptive, data-dependent audit rates.

Contributions.

1. **Formalization.** We cast unanimous LLM release as a selective-label mechanism and prove that zero-audit-probability strata have unidentifiable error rates, with disagreement-only escalation as the operative special case (§4).

2. **Measurement.** We measure joint false-accept rates across panel-composition axes on frozen judges and held-out tracks, separating *nominal* panel size from the *effective* number of independent judges (§5).

3. **Protocol.** We implement a randomized audit protocol with logged propensities, a positive audit floor, inverse-propensity estimation, and an anytime-valid Bennett–Ville certificate valid under adaptive audit allocation, with fail-closed logging and provenance (§6).

Table 1: Terminology and notation used throughout. Coined terms are marked †; all others are standard in the evaluation, causal-inference, or sequential-testing literatures.

Term / symbol	Meaning
LLM judge panel	A set of LLM judges J that each vote accept/reject on a candidate output.
Unanimous release	The gating rule: release automatically iff every judge accepts; escalate to a human otherwise. $Q_t{=}1$ if candidate t would be auto-released.
Silent consensus†	A released case that ordinary routing never labels: $Q_t{=}1$ and no independent outcome label is obtained.
Unsafe label Z_t	$Z_t{=}1$ if candidate t is unsafe under an independent adjudication (executable check or expert majority), fixed before any audit draw.
Joint false-accept rate (JFAR)	$q_J = \Pr\big(\bigcap_{j\in J}\{\text{judge } j \text{ accepts}\} \mid Z_t{=}1\big)$: the chance the whole panel approves an unsafe case.
Effective judge count	$D_{\text{eff}}(J) = \log \hat{q}_J \,/\, \big(\frac{1}{\lvert J\rvert}\sum_j \log \hat{q}_j\big)$: equals $\lvert J\rvert$ for independent equally erring judges, approaches 1 for clones.
Audit propensity π_t	Logged probability that released candidate t is randomly audited; $A_t \in \{0,1\}$ is the realized audit draw.
Release stratum S_t	A pre-specified subpopulation of release-eligible cases, defined by observable features only (e.g., a panel-confidence bucket)—the unit at which audit floors and error budgets are set, as in stratified sampling.
Positive support / audit floor	The requirement $\pi_t \geq \epsilon > 0$ in every release stratum (the positivity/overlap condition of causal inference).
Selective labels	Outcomes are observed only for cases the decision policy routes to labeling (Lakkaraju et al., 2017).
Anytime-valid bound	An upper confidence bound U_n with $\Pr(\forall n : H_n \leq U_n) \geq 1-\delta$, valid under continuous monitoring and optional stopping (Howard et al., 2021).
H_n, $\hat{H}_n$, U_n	Cumulative unsafe mass among released-eligible candidates; its inverse-propensity estimate; the anytime-valid upper bound.
Configuration epoch	A frozen (generator, panel, prompts, thresholds) configuration; any change starts a new epoch with fresh statistical state.

4. **Replays.** We replay audit policies on locked streams with pre-specified distribution shifts at matched audit budgets and quantify how much unsafe mass each policy leaves at zero audit probability (§7).

We do not claim that random audits make a system safe, that model-family diversity implies independence, or that executable labels replace human governance. The claim is narrower and, we believe, upstream of all of those: *a pipeline needs an independent observation channel with positive support before its deployed risk can be bounded at all*.

2 Related work

LLM judges and juries. Panels of smaller, heterogeneous judges can outperform a single large judge and reduce self-preference bias (Verga et al., 2024; Zheng et al., 2023), and panels can be selected by held-out competence, pairwise error correlation, or predicted instance-level reliability (Anonymous, 2026; Li et al., 2025). Unanimous LLM committees have been evaluated as conservative acceptance rules (Ullah & Serwadda, 2026). Errors nevertheless remain substantially correlated across architectures and providers (Kim et al., 2025), and repeated sampling of one model—as in self-consistency (Wang et al., 2023)—shares one model's blind spots by construction. Our problem begins *after* the panel is chosen: we study the labels its routing does and does not generate.

Scalable oversight and AI control. AI-control protocols combine untrusted strong models, trusted weaker monitors, editing, and scarce human labor under adversarial behavior (Greenblatt et al., 2024; Kale et al., 2025); monitorability depends on exposed process information (Guan et al., 2026); and LLM critics both catch and hallucinate bugs (McAleese et al., 2024). Our complement is statistical: even high-yield targeted escalation can assign *zero* probability to unanimous errors, and then nothing downstream can recover their rate.

Selective labels, audit sampling, and anytime-valid inference. Selective labels arise when decisions determine which outcomes are observable (Lakkaraju et al., 2017; Wei, 2021; Chen et al., 2025). Active testing for LLMs clips acquisition probabilities away from zero for unbiased evaluation (Berrada et al., 2025); proxy-assisted evaluation combines selective human audits, inverse-propensity corrections, and confidence sequences (Ao et al., 2026); deployment monitoring studies budgeted labeling and anytime-valid alarms (Ginart et al., 2022; Karampatziakis et al., 2021; Prinster et al., 2025; Timans et al., 2025); and randomized audit evidence has been tied to automation authorization (Deng, 2026). The statistical ingredients we use—positivity (Rosenbaum & Rubin, 1983), partial identification (Manski, 1990), inverse-propensity estimation (Horvitz & Thompson, 1952), martingale concentration and time-uniform bounds (Freedman, 1975; Ville, 1939; Howard et al., 2021; Ramdas et al., 2023)—are established. Our contribution is the formulation and the end-to-end contract: unanimous LLM release as a selective-label mechanism whose missing region the judge panel itself creates, plus the logging, floors, and provenance that keep the released population estimable.

3 Problem setup

Table 1 collects terminology and notation. At time t, a generator emits a candidate O_t with context X_t (governing policy, evidence, observable risk metadata). A panel of m judges emits votes $V_t = (V_{t1}, \ldots, V_{tm})$. Let $Q_t = 1$ when the frozen base policy would auto-release the candidate (here: unanimous accept, the strictest gate; majority-gate results appear in the artifact). Let $Z_t = 1$ mean the candidate is unsafe under an independent adjudication; Z_t is a potential label fixed relative to any audit draw.

For every release-eligible candidate (Q_t=1), a controller may log a randomized audit assignment $A_t \in \{0, 1\}$ with propensity

$$\pi_t \;=\; \Pr\big(A_t = 1 \mid \mathcal{F}_{t-1}, X_t, Q_t, S_t, G_t\big), \tag{1}$$

where S_t is a pre-specified observable stratum (e.g., a panel-confidence bucket) and G_t the configuration epoch. Audited cases are held before release until the label Z_t resolves (*pre-release holdback*; the label comes from an executable check or trusted expert process), and a pending or failed audit makes its stratum fail closed. The certificate targets cumulative unsafe mass among release-eligible candidates, $H_n = \sum_{t \le n: Q_t = 1} Z_t$; actual escaped harm (unsafe candidates that are actually released) is no larger under valid pre-release holdback.

Definition 1 (Silent consensus)**.** *A case is in silent consensus when $Q_t = 1$ and ordinary routing obtains no independent outcome label. A silent-consensus* error *additionally has $Z_t = 1$.*

More judges do not automatically mean more protection, so we measure joint failure and effective independence separately. On independently labeled unsafe cases, the **joint false-accept rate** of panel J is $q_J = \Pr\big(\bigcap_{j \in J}\{\text{judge } j \text{ accepts}\} \mid Z_t = 1\big)$, and the **effective judge count** D_{eff} (Table 1) summarizes how far a panel's joint behavior is from independence: three clones give $D_{\text{eff}} \approx 1$; three independent equally erring judges give $D_{\text{eff}} = 3$.

4 Why disagreement-only auditing cannot certify releases

The observed pipeline log contains $(X_t, Q_t, S_t, G_t, A_t, A_t Z_t)$: outcomes are visible only for audited cases. The result below claims no new mathematics; its roots are classical. Rubin (1976) showed that outcome inference fails when missingness depends on the decision process unless that dependence is ignorable or the missingness probability is known; Rosenbaum & Rubin (1983) made positivity (overlap) a pillar of propensity-based inference; and the $[0, 1]$ identified set below is exactly Manski's worst-case bound (Manski, 1990). What is new is the missingness mechanism: labels are deleted by an LLM judge panel, precisely on the cases it is most confident about. The proof is elementary; we state it because its corollary is the operative design constraint.

Theorem 1 (Unaudited strata are unidentifiable)**.** *Consider a release-eligible stratum s with positive population mass. If $\pi_t = 0$ almost surely for all cases in s, then the unsafe rate of s is not nonparametrically point identified from the pipeline log. Without additional outcome assumptions or an external label channel, the identified set is* $[0, 1]$.

Proof. In stratum s, $A_t = 0$, so $A_t Z_t \equiv 0$ regardless of Z_t. Construct two worlds with identical laws for all observed variables and all audited outcomes outside s: set $Z_t = 0$ for every case in s in the first and $Z_t = 1$ in the second. The observed-data distributions coincide while the target rate is 0 versus 1; mixtures give every intermediate value. □

Corollary 1 (Disagreement blind spot)**.** *If audits are triggered only by judge disagreement and the base policy releases unanimous approvals, then every unanimous unsafe candidate has zero audit probability. Its unsafe rate is not identified by any quantity of disagreement labels.*

Disagreement-only auditing is not merely biased; it leaves the released region *unbounded*. Calibration, confidence thresholds, and disagreement statistics are all functions of the same log and cannot help. The constructive counterpart is standard: if $\pi_t \geq \epsilon > 0$ is logged, the inverse-propensity (Horvitz–Thompson) estimator recovers the released risk, even when audit rates adapt to past outcomes.

Proposition 1 (Unbiasedness under adaptive audit rates)**.** *Let $\hat{H}_n = \sum_{t \leq n} A_t Z_t / \pi_t$ over release-eligible candidates within a single stratum and configuration epoch. If π_t is predictable (measurable with respect to decision-time information) and the audit draw is independent of Z_t given that information, then $\hat{H}_n - H_n$ is a martingale, even when future propensities depend on all past audited outcomes.*

Proof. Conditional on decision-time information and Z_t, $\mathbb{E}[A_t Z_t / \pi_t] = Z_t$; summing the martingale differences proves the claim. □

An anytime-valid certificate. Deployment monitoring is continuous, so a fixed-n confidence interval is the wrong object: the bound must hold at every prefix. Let $W_t = Z_t - A_t Z_t / \pi_t$. Given decision-time information, W_t has mean zero, $W_t \leq 1$, and conditional second moment at most $(1 - \pi_t)/\pi_t$. With $V_n = \sum_{t \leq n} (1 - \pi_t)/\pi_t$ and $\phi(\lambda) = e^\lambda - 1 - \lambda$, the Bennett moment bound makes $\exp\{\lambda(H_n - \hat{H}_n) - \phi(\lambda) V_n\}$ a nonnegative supermartingale for each fixed $\lambda > 0$ (Freedman, 1975). Fixing a grid $\Lambda = \{2^{-6}, \ldots, 2^2\}$ before any outcome and spending $\delta/9$ per grid point, Ville's inequality (Ville, 1939) gives the explicit, time-uniform certificate

$$U_n \;=\; \min\Big\{ n,\; \hat{H}_n + \min_{\lambda \in \Lambda} \frac{\log(9/\delta) + \phi(\lambda) V_n}{\lambda} \Big\}, \qquad \Pr\big(\forall n : H_n \leq U_n\big) \geq 1 - \delta. \tag{2}$$

The construction admits adaptive predictable propensities. Two consequences shape everything downstream. First, the bound's width scales like $\sqrt{V_n \log(1/\delta)}$, and low audit rates inflate V_n: *tightness is purchased with audit mass, not with better mathematics*. Second, Theorem 1 says the certificate cannot exist at all for strata with $\pi_t = 0$—which is exactly where disagreement-only routing puts every unanimous error.

5 Measuring silent consensus

5.1 Setup: frozen judges, held-out tracks, independent labels

Tracks. We evaluate on three source-disjoint tracks whose ground truth does not come from any LLM judge: **PolicyDelta** (versioned enterprise-policy cases with executable, typed labels; stressors: clause revision, no-ops, schema shift), **ReasonBench-logical** (claims over independently generated logical worlds, i.e., rule sets, with executable labels; stressors: reordering, inversion, unseen composition), and **HealthBench** (Arora et al., 2025; medical responses with physician-majority binary rubric labels; stressors: unseen expert criteria). Calibration, development, held-out test ("locked": frozen and untouched until final evaluation), and shift splits are disjoint at the source-unit level (clause, world, conversation). Unsafe prevalences are designed for stress testing (e.g., each ReasonBench claim is paired with a corrupted variant), so no number below is a deployment rate. Locked-split sizes: PolicyDelta 86 cases (48 unsafe, 6 clause clusters), ReasonBench 384 (192 unsafe, 96 worlds), HealthBench 521 (84 unsafe, 146 conversation clusters).

Judges. The frozen panel (fixed before any test-set inference) contains commit-pinned open-weight judges spanning controlled composition axes: a same-family size triple (Qwen3-8B/4B/1.7B); four cross-family judges (Mistral-7B-Instruct-v0.3, OLMo-2-7B-Instruct, Phi-4-mini, GLM-4-9B); one

Table 2: Joint false-accept rate (JFAR) on held-out (locked) test units under unanimous release (confirmatory), with the fraction of *safe* cases each gate escalates to humans. Safe n: PolicyDelta 38, ReasonBench 192, HealthBench 437. †One-sided 95% upper bounds for zero-event cells at case level / source-cluster level (PolicyDelta has only 6 locked clause clusters; ReasonBench 96 worlds); both are shown because renderings within a cluster are not independent. Zero observed joint failures are never zero risk.

Track (locked)	Panel	JFAR	95% CI	Joint/unsafe	Safe escal.
PolicyDelta	same-family (Qwen3 8B/4B/1.7B)	0.0%	[0, 6.2 / 39.3]†	0/48	63%
PolicyDelta	cross-family 5	0.0%	[0, 6.2 / 39.3]†	0/48	97%
ReasonBench	same-family	1.0%	[0, 2.6]	2/192	99%
ReasonBench	cross-family 3	0.0%	[0, 1.6 / 3.1]†	0/192	99%
ReasonBench	cross-family 5 + executor	0.0%	[0, 1.6 / 3.1]†	0/192	100%
HealthBench	same-family	**42.9%**	[31.7, 54.6]	36/84	27%
HealthBench	cross-family 3	22.6%	[13.8, 32.3]	19/84	59%
HealthBench	cross-family 5	2.4%	[0, 5.9]	2/84	83%

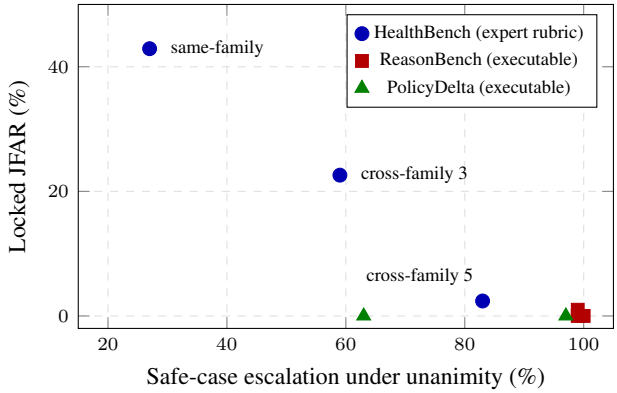


Figure 2: Silent consensus trades off against automation. Each point is one locked panel from Table 2: cross-family panels lower the observed joint false-accept rate but escalate most safe work to humans. Points are observed contrasts, not a continuous intervention—membership, capability, and format compliance change together—so this is a correlation, not a causal effect of diversity. Zero-JFAR panels retain nonzero upper bounds.

meaning-preserving prompt paraphrase and two temperature-0.7 samples of the Qwen3-8B anchor; and a deterministic symbolic executor for the logical track that forward-chains the judge-visible rules and abstains (fail-closed) on any input it cannot fully parse. Every output is bound to model revision, prompt hash, and decoding settings. Invalid, truncated, or abstaining responses count as non-accepts; headline cells are unchanged under strict parsing (Appendix A).

Evidence grading. Locked-split panel results are confirmatory under our internal pre-specified protocol (frozen before outcomes; not an external preregistration). Composition-axis results computed on calibration units are pre-specified but *descriptive*: the same units informed calibration choices. Confidence intervals are cluster bootstraps at the source-unit level; zero-event cells report one-sided rule-of-three upper bounds ($\approx 3/n$ at the relevant unit). Two independent technical audits reproduced the headline numbers from raw artifacts, and an append-only deviation log records every amendment.

5.2 Result: Silent consensus is large where judgment is subjective

Table 2 and Figure 2 give the central measurement. On the physician-rubric track, where labels encode expert judgment rather than an executable check, the same-family panel unanimously ap-

Table 3: Effective independence by panel-composition axis on calibration units (pre-specified, descriptive; three nominal judges except *two, all sequences anchored at Qwen3-8B). D_{eff} (computed with add-half smoothing) equals the nominal count for independent equally erring judges and approaches 1 for clones. ‡Zero-event one-sided upper bounds at case / clause-cluster level.

Track (calibration)	Panel	JFAR	95% CI	Joint/unsafe	D_{eff}
PolicyDelta	same model, 3 samples	4.5%	[0.8, 8.9]	12/268	1.07
PolicyDelta	prompt paraphrase pair*	4.9%	[0.8, 10.3]	13/268	1.15
PolicyDelta	same family, 3 sizes	2.2%	[0.4, 5.2]	6/268	1.77
PolicyDelta	cross-family 3	0.0%	[0, 1.1 / 11.1]‡	0/268	2.23
ReasonBench	same model, 3 samples	**39.0%**	[30.0, 49.0]	39/100	1.37
ReasonBench	same family, 3 sizes	1.0%	[0, 4.0]	1/100	2.26
ReasonBench	cross-family 3	3.0%	[0, 6.0]	3/100	2.40
HealthBench	same model, 3 samples	**43.1%**	[30.8, 55.6]	22/51	1.26
HealthBench	same family, 3 sizes	31.4%	[18.5, 43.9]	16/51	2.74
HealthBench	cross-family 3	23.5%	[12.0, 35.3]	12/51	2.15

proved **42.9%** (36/84) of criterion-violating responses while automating 73% of safe ones. Growing the panel to five model families cut the observed JFAR to 2.4%—but escalated 83% of *safe* cases to humans, nearly erasing the automation the gate exists to provide. Across the tested sequence, lower JFAR coincided with higher safe-case escalation ($27\% \to 59\% \to 83\%$), and joint failures persisted even at the largest panel. Membership, capability, and format behavior change together across panels, so the contrast is descriptive, not a causal effect of diversity. On the executable tracks, joint failures were rare on locked units, but the zero cells are small-sample zeros: with only 6 locked clause clusters, PolicyDelta's cluster-level upper bound is 39.3%. Observed joint failures exceed the corresponding independence product several-fold, and blind-spot location shifted between surface-form families in ways family metadata did not predict (Appendix B).

5.3 Result: more judges do not mean more independent judges

Table 3 asks whether adding judges actually adds protection. Three temperature-0.7 samples of one 8B model—the panel a self-consistency-style vote implicitly builds (Wang et al., 2023)—behaved like **1.1–1.4 effective judges** and jointly approved 39.0% of corrupted reasoning claims and 43.1% of unsafe medical responses. A meaning-preserving prompt paraphrase added almost as little ($D_{\text{eff}} = 1.15$). Same-family checkpoints reached 1.8–2.7 effective judges out of three; cross-family panels reached only 2.15–2.40. Under the tested conditions, counting judges did not predict how independent they were: *a three-judge panel can be one judge wearing three hats*. The deterministic executor is the opposite limiting case: within its closed grammar it is exact (0% false accepts, 100% safe accepts on locked ReasonBench), but under linguistic restatement it abstains on every case—its coverage boundary is itself a shift family, so mechanistic verification cannot anchor assurance alone (Appendix A).

6 A randomized audit protocol with an anytime-valid certificate

Theorem 1 dictates the design: every release-eligible stratum must carry a logged, positive audit probability. The protocol composes a mandatory floor with an optional targeted component,

$$\pi_t \;=\; 1 - (1 - \epsilon_{S_t,G_t})\,(1 - \pi_t^{\text{target}}), \qquad \epsilon_{S_t,G_t} > 0, \tag{3}$$

so targeting can concentrate audits but can never zero out a stratum. A trusted controller owns the propensity plan; callers cannot supply case-level probabilities.

Logging and fail-closed operation. The stratum, executed propensity, keyed pseudorandom draw (with key commitment), assignment, payload and decision hashes, and configuration digest are logged *before* outcome access; audited candidates are held pre-release until the label resolves; a pending, failed, or mismatched audit fails its stratum closed. Controller state is write-ahead journaled to a hash-chained log, so restarts replay to bit-identical state. Estimation uses the exact im-

Table 4: The identifiability boundary in practice (corrected replays, exploratory): post-shift unsafe releases assigned exactly zero audit probability by each policy (273 events = 91 unique unsafe cases × 3 arrival orders). The targeted arms are retrospective oracles with the whole stream's score ranking—an upper bound on any online threshold rule. Mass at $\pi = 0$ cannot be estimated from that policy's log by any downstream method (Theorem 1).

Audit policy	**Unsafe mass at $\pi = 0$**	**Share**	**Positive support?**
Disagreement-only	273/273	100%	no
Retrospective confidence-targeted (oracle)	187/273	68.5%	no
Retrospective risk-targeted (oracle)	187/273	68.5%	no
ϵ-greedy (4% floor)	0/273	0%	yes
Uniform random	0/273	0%	yes
Stratified random (weighted)	0/273	0%	yes
Full protocol (2% floor + targeting)	0/273	0%	yes

plemented propensities in $\hat{H}_n$ and the certificate U_n of equation 2; per stratum, an e-process (a sequential test that stays valid under optional stopping; Ramdas et al., 2023) raises an alarm against a pre-registered tolerated rate, and any changed component (model, prompt, threshold) starts a new configuration epoch whose evidence is not inherited. A prospective budget gate—authorize the next exposure only while $U_n + 1$ fits a linear harm allowance—is implemented and unit-tested but not evaluated in closed loop here; validity assumptions (label independence from the panel, executed-equals-logged propensities, completed audits) are stated in Appendix D.

Verification. The implemented certificate showed no coverage violation in 4,800 synthetic trials spanning four harm regimes and three adaptive propensity schedules, nor at any finalized prefix of 27 journaled replay traces; realized widths were conservative (final upper/true ratio: median 4.1, range 3.1–19.0). These checks stress-test the implementation, not the theorem; tighter certificates require more independent audits, not weaker mathematics.

7 Replayed deployments: what a positive audit floor buys

Setup. We replay audit policies on locked streams built from the held-out and shift splits: each stream is a stationary segment followed by one pre-specified distribution shift (policy-operator substitution, linguistic restatement, unseen composition/deeper proofs, or held-out expert criteria), with the release panel per track chosen by a frozen calibration-only screen (Appendix A); gold labels are hidden online and revealed only as each policy requests them. Policies: disagreement-only; uniform random; stratified random (2:1 weighting toward the higher-risk stratum); ϵ-greedy (4% floor); two *retrospective* confidence/risk-targeted oracles that rank the entire future stream by observable score (unattainable online; panel confidences cluster on a few values); the full protocol of §6 (2% floor, 25% target); a floor-only ablation; and a hidden-label oracle as a diagnostic ceiling. Positive-support arms are approximately rate-matched (realized 27.8–28.4% of release-eligible arrivals, within 2%); disagreement-only audits nothing eligible by construction and is not cost-matched. An independent audit found design defects in the first execution (a miscalibrated alarm null, a degenerate baseline, unrecorded parameter drift); every number below is from the corrected rerun and labeled **exploratory**. The scenario (six shift scenarios × three arrival orders) is the unit of analysis.

Targeted auditing fails exactly where it is needed. Table 4 operationalizes Theorem 1. The confidence oracle—shown every score in the stream in advance, but no labels—still left **68.5%** of post-shift unsafe releases at zero audit probability: shared blind spots sit where the panel is *confidently* wrong, and the targeting score comes from the same panel whose blindness is being audited. Disagreement-only routing left 100%. Every policy with a positive random floor, even at 2%, kept every one of these cases auditable. Better scores cannot close this gap; only randomization can.

A random floor makes risk estimable and blocks some unsafe outputs; it does not crown a best policy. Across scenarios, every positive-support arm cut post-shift escaped unsafe releases from 91.0 (disagreement-only) to 62.3–68.0 cases (averaged over the three arrival orders), blocking 23.0–28.7

during pre-release holdback (25–32% in aggregate), at a realized ≈28% audit rate; safe throughput (the fraction of safe cases released without human review) was 0.67–0.69 versus 1.00. This contrast bundles positive support, audit volume, and holdback (disagreement-only is not cost-matched), so it shows the value of auditing at all, not of any one policy. Two pre-specified comparisons were *inconclusive*; we build no claim on them: (i) the full protocol versus stratified random (the strongest development baseline) showed a paired mean difference of 0.0 post-shift escapes (sign-permutation 95% CI $[-0.6, +0.6]$, n=6, two-sided p=1.0), with the floor-only ablation in the same range; (ii) under correctly calibrated per-cell alarm thresholds, *no* shift alarm fired in any of the 18 shifted traces—post-shift segments of 28–125 release-eligible arrivals at a 28% audit rate carry too little audited harm to cross an honest anytime threshold; detection needs longer streams or more audit mass, and the first execution's alarms are retracted (full accounting: Appendix C).

8 Discussion

What is established. Theorem 1 is an observability condition: zero audit probability means the log cannot bound the unsafe rate. The measurement study shows joint false acceptance can be large (36/84 physician-labeled violations for a same-family panel) and zero-event cells retain nonzero cluster-level upper bounds; certificate coverage held in every synthetic and replay check (§6).

What remains exploratory or descriptive. The 68.5% zero-support share, blocked-case counts, null policy ranking, and absent shift detection all come from six corrected stress scenarios; none supports superiority, equivalence, or closed-loop control. The same-model sampling result is pre-specified but computed on calibration units (§5) and awaits locked-split replication.

Implications for practice. Three lessons travel. (1) *Do not read silence as safety*: a pipeline that labels only disagreements measures its own routing, not its risk. (2) *Do not count judges; count effective judges*: three samples of one model, or a prompt paraphrase, act as roughly one judge; even cross-family panels fall short of independence. (3) *Reserve a small random audit floor and log its probabilities*: it is cheap (2–4% sufficed here), the precondition for any risk estimate over released outputs, and it composes with targeted escalation. It does not by itself deliver precision (widths $\sim$4$\times$ truth here), fast detection, or accountability; those cost more audit mass, longer horizons, and named human owners.

9 Limitations

Our tracks are controlled stress tests with designed unsafe prevalences; escaped-event counts describe policy behavior under those designs, not deployment harm rates. Positive support makes risk estimable, not the adjudicator correct: executable labels are exact only relative to their declared semantics, expert labels carry disagreement, and the guarantee assumes valid, promptly resolved labels independent of the panel; delay, nonresponse, and label noise fall outside the theorem, and the certificate bounds cumulative unsafe mass, not per-candidate risk or severity. Judge lineage is a risk feature, not proof of (in)dependence (training-data overlap is unobservable). The replays exercise certificates and alarms, not the prospective budget gate, and do not test adversarial prompt injection. Time-uniform bounds are data hungry: too few audits mean loose certificates; our alarms detected none of the injected shifts. Finally, this work does not decide whether a policy is legitimate, fair, or lawful, nor replace appeals and accountable human ownership; its purpose is to stop an evaluation pipeline from treating the absence of labels—an absence it created—as evidence of safety.

10 Conclusion

Unanimous LLM release creates a selective-label problem: the panel decides which outcomes are observed, and the unobserved region is exactly what it releases. We proved that region's error rate unidentifiable without positive audit probability, and showed—in measurement and in replay—that shared blind spots are large and only a small logged random audit floor keeps released risk estimable. The recommendation: for every released population, log a positive audit probability and obtain valid independent labels; that makes deployed risk measurable, and everything else is purchased on top.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods (experimental and analysis code). We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or write proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat datasets, or interpret results; translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability. We have reviewed all AI-assisted work: all AI-assisted code and text edits were reviewed by all authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All numbers regenerate from recorded run packages and scripts in the accompanying artifact (anonymized for review): commit-pinned judge checkpoints with per-file SHA-256 manifests, frozen prompts and decoding settings, split digests, the frozen internal protocol with its append-only deviation log (including the corrections in §7), journaled replay traces with hash-chained provenance, and the exact implemented audit propensities used in every estimator. Validity assumptions appear in §4 and Appendix D; evidence-grade labels (confirmatory/descriptive/exploratory) are stated inline where each result is reported.

REFERENCES

Anonymous. Correlation-aware ranking and filtering for heterogeneous small-model judge ensembles. Anonymous concurrent submission under double-blind review, 2026.

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.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning with selectively labeled data from multiple decision-makers. In *International Conference on Machine Learning*, pp. 8480–8519, 2025.

Jie Deng. When can fraud operations authorize automation? a decision-support framework for fresh audit evidence and review workload. *arXiv preprint arXiv:2608.08577*, 2026.

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

Tony Ginart, Martin Jinye Zhang, and James Zou. MLDemon: Deployment monitoring for machine learning systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 3962–3997, 2022.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *International Conference on Machine Learning*, 2024.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring monitorability. In *International Conference on Machine Learning*, 2026.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

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.

Neil Kale, Chen Bo Calvin Zhang, Kevin Zhu, Ankit Aich, Paula Rodriguez, Scale Red Team, Christina Q. Knight, and Zifan Wang. Reliable weak-to-strong monitoring of LLM agents. *arXiv preprint arXiv:2508.19461*, 2025.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *International Conference on Machine Learning*, pp. 5301–5310, 2021.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *International Conference on Machine Learning*, pp. 30038–30066, 2025.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Xiaochuan Li, Ke Wang, Girija Gouda, Shubham Choudhary, Yaqun Wang, Linwei Hu, Joel Vaughan, and Freddy Lecue. Who judges the judge? LLM jury-on-demand: Building trustworthy LLM evaluation systems. *arXiv preprint arXiv:2512.01786*, 2025.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM critics help catch LLM bugs. *arXiv preprint arXiv:2407.00215*, 2024.

Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *International Conference on Machine Learning*, pp. 49830–49859, 2025.

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

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A. Naesseth. On continuous monitoring of risk violations under unknown shift. In *Conference on Uncertainty in Artificial Intelligence*, pp. 4204–4226, 2025.

Muhammad Aziz Ullah and Abdul Serwadda. Vibe coding on trial: Operating characteristics of unanimous LLM juries. *arXiv preprint arXiv:2602.18492*, 2026.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, 1939.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International Conference on Machine Learning*, pp. 11035–11046, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A PER-JUDGE MARGINAL BEHAVIOR AND PARSING SENSITIVITY

Table 5: Locked-unit marginal behavior per judge: false-accept rate on unsafe cases (FAR, %) and accept rate on safe cases (%). Invalid, truncated, or abstaining responses count as non-accepts.

	PolicyDelta		**ReasonBench**		**HealthBench**	
Judge	FAR	Accept	FAR	Accept	FAR	Accept
Qwen3-8B	0.0	92.1	51.6	86.5	60.7	87.9
Qwen3-4B	0.0	89.5	50.0	65.1	79.8	95.9
Qwen3-1.7B	6.2	36.8	1.0	1.0	73.8	85.6
Mistral-7B-Instruct-v0.3	0.0	47.4	71.4	87.0	36.9	59.3
OLMo-2-7B-Instruct	2.1	65.8	0.0	1.0	69.0	79.6
Phi-4-mini	0.0	81.6	6.8	4.2	61.9	78.5
GLM-4-9B	6.2	5.3	0.0	0.0	32.1	42.3
Deterministic executor	abstains		0.0	100.0	abstains	

Judges divide into three recurring behaviors. *High-acceptance* judges accept most safe work and much unsafe work (Qwen3-8B false-accepts 51.6% of corrupted logical claims; Mistral 71.4%). *Low-coverage* judges reject or fail to format nearly everything (OLMo-2 and Qwen3-1.7B on ReasonBench; GLM-4 broadly—its verbosity overruns the frozen 160-token budget); a unanimity panel containing one has an empty release region by construction, which vacates the estimand instead of measuring it, motivating the calibration-only deployability screen used for the replay panels (a judge qualifies if it accepts at least half of calibration-safe cases; top three by safe-accept rate serve). The deterministic executor is exact on its closed grammar and silent everywhere else. Judge competence is task-local: family, size, and benchmark rank did not transfer across judgment types.

Parsing sensitivity. Phi-4-mini and GLM-4-9B wrap valid JSON in markdown fences; a frozen extraction rule recovers a decision only when exactly one schema-valid object is present. The headline JFAR cells are robust to this amendment: under strict parsing alone, the HealthBench same-family cell stays 36/84 and cross-family-3 stays 19/84. The cross-family-5 residual (2/84) appears only after format recovery lets two otherwise all-invalid judges vote; without extraction, their blanket invalidity would have produced a vacuous zero. Part of what a unanimity gate blocks is malformed output, so cross-family panels owe part of their escalation cost to format failure rather than judgment.

B BLIND-SPOT STRUCTURE ACROSS SHIFT FAMILIES

Splitting PolicyDelta's shift families on locked units: under linguistic restatement no panel showed a joint failure (0/46); under operator substitution the five-family panel false-accepted 1/16 (6.3%) while the same-family panel showed none. This reverses the calibration-split ordering, on a small denominator (3 clauses). In these samples, model-family metadata did not predict where shared blind spots sat. Where joint failures exist, they exceed a track- and transformation-conditioned independence product several-fold; we report this as a descriptive dependence ratio, not a causal decomposition, because within-cell item difficulty is not separately controlled. On locked ReasonBench, the cross-family panels already show no joint failure, so the deterministic executor's marginal value there is the bound it can *certify* rather than errors it removes; on the same-family panel's two residual failures it rejects both. Under linguistic restatement it abstains on every case: certificates need executors where they exist and randomized audit labels where they do not.

C CORRECTED REPLAY ACCOUNTING

The pre-specified paired comparison of the full protocol against this baseline gives per-scenario post-shift differences $\{+1.7, 0, -1, 0, -0.3, -0.3\}$: mean 0.0 (sign-permutation 95% CI $[-0.6, +0.6]$, two-sided $p = 1.0$, $n = 6$), with a macro-average safe-throughput difference of $+0.2$ points (per-scenario range -5.7 to $+6.1$), so a -5-point non-inferiority margin holds on average but is not established per scenario. The set is not powered for equivalence, non-inferiority, or absence of

Table 6: Corrected replay, scenario-level accounting (exploratory). Post-shift columns use the 91 unique post-shift unsafe cases as denominator (blocked = 91 − escaped; no alarms fired). Whole-stream columns sum per-scenario means over all nine scenarios including stationary segments and null streams, and are not comparable to the 91-case denominator. Audit totals of positive-support arms are approximately rate-matched within 2% (realized 27.8–28.4% of release-eligible arrivals); disagreement-only and the hidden-label oracle audit none or exactly the unsafe cases by construction and are not cost-matched. Safe throughput is a scenario-equal-weight macro-average; pooled over safe cases it is 0.72 for every positive-support arm. Differences among positive-support arms are within arrival-order noise.

	Post-shift (of 91)		**Whole-stream (9 scen.)**			
Arm	Escaped	Blocked	Escaped	Blocked	**Audits**	**Safe thr.**
Full protocol (2% floor + targeting)	63.3	27.7	149.0	62.0	458	0.69
Floor-only (no targeting)	66.7	24.3	153.7	57.3	454	0.69
Stratified random (best dev baseline)	63.3	27.7	146.7	64.3	459	0.68
Uniform random	68.0	23.0	157.3	53.7	450	0.67
ϵ-greedy	62.7	28.3	147.0	64.0	459	0.68
Retro. confidence-targeted (oracle)	62.3	28.7	146.0	65.0	458	0.68
Retro. risk-targeted (oracle)	62.3	28.7	146.0	65.0	458	0.68
Disagreement-only	91.0	0	211.0	0	0	1.00
Hidden-label oracle (diagnostic)	0	91.0	0	211.0	211	1.00

information in the panel-confidence signal; the structural result of Table 4 is what the design does establish. Audit-controller settings: 2% floor, 25% global target, 32-arrival reallocation blocks, 8-candidate per-cell burn-in, fixed from development-stream detection power. Alarm thresholds are one-sided Wilson 95% upper bounds of each monitored (track, confidence-stratum) cell's stationary eligible-unsafe rate on development data; a cell with no development data receives a conservative 0.5 threshold and is reported as uncalibrated (one cell, with seven development observations, is close to inert). Two earlier alarm calibrations (a flat 5% null, then track-pooled bounds) were each false on at least one monitored cell; their alarms are retracted.

D PROTOCOL VALIDITY ASSUMPTIONS AND ADDITIONAL DETAILS

The certificate of equation 2 is design-based and rests on the following stated assumptions. (i) Conditional on the complete pre-draw filtration, the 53-bit keyed draw is independent and discrete-uniform; the HMAC implementation satisfies this only under an ideal-PRF model with a secret, one-time random key. Production deployments should generate the key in a secret store, publish only its commitment, and rotate it with configuration epochs; the fixed key in our replays is a public reproducibility seed and provides no adversarial unpredictability. The implementation draws $U = K/2^{53}$ and therefore uses the exact implemented propensity $\lceil 2^{53} p_t \rceil / 2^{53}$ in every estimator and test. (ii) The audited label is valid for the original candidate and independent of the judge panel; there is no interference between candidates. (iii) The recorded propensity equals the executed one; the decision-time schema excludes outcome-like fields, and assignment, execution, and outcome must carry matching configuration digests. (iv) Every selected audit completes (enforced by fail-closed pre-release holdback); with nonresponse, the effective observation probability is the product of assignment and completion propensities, and treating completion as ignorable is a further assumption—which is why a pending or failed selected audit forces the trivial certificate rather than a silent substitution. (v) Audit-mass reallocation happens only between completed blocks, using labels finalized before the block starts, with the reallocation rule bound into the epoch's configuration digest. (vi) Outcome-submitter authentication is a deployment assumption, not an enforced property of the reference implementation. Each run registers finite strata and at most eight epochs, allocating error budgets so $\sum \delta_{s,g} \leq 0.05$; exceeding the registry fails closed rather than resetting evidence. Label blinding in the replays is interface-level within one process; a hardened deployment would isolate the label store.