Robust Doubly-Efficient Debate with Correlated, Contaminated Human Judgment

Liz Lemma Future Detective

January 23, 2026

Abstract

Doubly-efficient debate provides a complexity-theoretic route to scalable oversight: two polynomial-time provers can convince a much cheaper verifier using only a constant number of oracle queries, even when the oracle represents black-box human judgment. Existing stochastic results crucially rely on i.i.d. oracle samples and Chernoff bounds. In modern 2026 oversight pipelines, however, human judgments are correlated (shared rubrics, common-mode bias, evaluator drift) and partially adversarial (brigading, compromised contractors, reward-model feedback loops). We extend the stochastic doubly-efficient debate framework by replacing the verifier's i.i.d. sampling step with robust mean estimation under (i) Huber contamination and (ii) controlled dependence quantified by a design-effect parameter. We show that a median-of-means (or trimmed) audit achieves the same qualitative guarantee as the original protocol: constant human query complexity independent of the length of the underlying computation, with completeness–soundness gap degrading gracefully in the contamination rate. We provide a clean parameterization that separates (a) task sensitivity (the Lipschitz constant K) from (b) evaluation-system reliability (contamination ε and dependence χ), and discuss empirical calibration via clustered-rater simulations as a guide for institutional panel design.

Table of Contents

1. 1. Introduction: why i.i.d. human judgments are the wrong abstraction in 2026; overview of robust debate as an auditing primitive.

2. 2. Background and baseline: recap of stochastic doubly-efficient debate (protocol structure, where Chernoff enters, role of K-Lipschitzness).

3. 3. A clean evaluator model: contamination + dependence. Define $\widetilde{O}$, ε, and the design-effect χ. Discuss interpretable special cases (clustered panels, drifting raters).

4. 4. Robust estimation toolkit: median-of-means / trimmed mean for Bernoulli means under Huber contamination and sub-Gaussian dependence inflation. State finite-sample bounds with explicit constants.

5. 5. Robustified debate protocol: modify the abort-audit step; specify thresholds and sample sizes m as functions of K, χ, δ.

6. 6. Main theorem (completeness/soundness): prove robust debate maintains a constant gap with $O(\chi K^2 \log(1/\delta))$ human ratings independent of T; quantify degradation in ε.

7. 7. Tightness and limits: impossibility when ε is too large; discussion of when dependence inflation χ must scale (and why constant-cost breaks).

8. 8. Practical calibration and simulations (flagged as numerical): clustered-rater generative models; mapping observed intra-class correlation to χ; sensitivity of audit sizes and error.

9. 9. Implications for oversight design: panel procurement (diversity, independence), robust aggregation standards, and integration with reward models; open problems.

1 Introduction: why i.i.d. human judgments are the wrong abstraction in 2026; robust debate as an auditing primitive

In much of the theoretical literature on human oversight, a convenient abstraction is that we can sample an i.i.d. panel of "judges" whose answers are unbiased draws from some fixed distribution. This assumption is appealing because it lets us import concentration inequalities essentially for free: if each rating is an independent coin flip with mean p, then a sample average estimates p to accuracy $\tilde{O}(m^{-1/2})$, and the rest of the analysis becomes bookkeeping. Our claim is that this abstraction is miscalibrated for the oversight regimes that matter in 2026. The most salient deviations from i.i.d. are not small perturbations; they are structural features of the socio-technical pipeline that produces "human labels" for complex, high-stakes model behavior.

First, independence is routinely violated. In deployed workflows, raters are trained on shared rubrics; they consume overlapping context; they influence each other through discussion boards, exemplars, and institutional defaults; and they are often recruited from correlated populations. Even absent explicit coordination, shared priors and shared failure modes generate intra-class correlation: if a query is phrased in a misleading way, many raters will err together. From a statistical perspective, the effective sample size can be far smaller than the nominal panel size. Moreover, correlation is not merely an inconvenience; it can be endogenous to the adversary. If a model (or a strategic prover) can choose which questions to ask and how to frame them, it can steer attention toward ambiguity where human judgments are most correlated and least reliable, making naive audits systematically overconfident.

Second, unbiasedness is routinely violated in ways that are well modeled by contamination rather than by mean-zero noise. Some fraction of ratings may be careless, rushed, or systematically biased by ideological or financial incentives; some may be adversarial (e.g. a compromised rater pool, prompt-injected annotator instructions, or a "brigading" event); and some may simply reflect distribution shift, where the rubric does not apply. If we insist on modeling all of these as i.i.d. noise, we implicitly assume that errors cancel out in expectation. In practice, the oversight failure modes we fear are precisely those where errors do *not* cancel: a small coalition can drive the estimate by a constant amount, or a shared misunderstanding can flip the sign of the decision. This motivates robust statistics: we would like estimators whose performance degrades gracefully with an explicit contamination rate ε, rather than collapsing once a small number of samples are adversarial.

These observations force a reframing. The object we actually need from

humans in a scalable oversight stack is not an omniscient ground-truth oracle, and not even a perfectly calibrated probability estimate. Instead, we need a bounded-cost *auditing primitive*: a procedure that, given a single contested subclaim in a larger computational argument, can probabilistically distinguish "the prover's claim is meaningfully wrong" from "the prover's claim is approximately right," despite the fact that the human evidence is noisy, correlated, and partially corrupted. This is the gap between labeling and verification. Labeling asks humans to answer the whole question; verification asks humans to test a targeted vulnerability in a purported reasoning trace.

Robust debate is attractive because it turns verification into a game with a single point of empirical contact. The high-level idea is to force a powerful agent A that claims "the computation is correct" to expose intermediate probabilistic assertions, while an opposing agent B is incentivized to locate a step where A's assertion diverges from reality. The verifier V does not need to inspect the full computation; it needs only to conduct a one-shot audit at the disputed step. Economically, this matches how oversight is actually budgeted: human attention is the scarce resource, and we can afford to spend it only once per task (or at most a constant number of times), even as the model's internal reasoning trace becomes arbitrarily long.

The obstacle is that the classical analysis of such protocols typically relies on Chernoff-style concentration at the audit step: if the verifier draws m independent samples from an oracle and takes the empirical mean, then with high probability the estimate is close to the truth. Once we leave the i.i.d. world, that inference no longer holds. Correlation inflates variance, and contamination introduces bias. If we continue to use the empirical mean anyway, a strategic prover can exploit the gap: it can craft disputed steps whose true conditional probability p sits near the verifier's acceptance threshold and then rely on the audit's miscalibration to slip through. This is a safety issue, not a technicality, because the entire point of the protocol is to make it expensive for the prover to lie; a fragile audit statistic reopens the attack surface.

Our approach is to explicitly separate two roles that are often conflated: (i) the debate protocol, which is a complexity-theoretic reduction that says "if you can accurately check these local probability claims, then you can verify the global computation"; and (ii) the human aggregation rule used in the audit, which is a statistical component that says "given m noisy ratings, we can estimate the local probability claim to tolerance τ." Once separated, we can robustify (ii) without changing the conceptual structure of (i). Concretely, instead of treating the human panel as i.i.d. draws, we model it as a mixture of an uncorrupted process and an adversarial process, with a dependence parameter capturing variance inflation. We then plug in a robust mean estimator (e.g. median-of-means) whose deviation guarantees degrade as $\tilde{O}(\sqrt{\chi/m} + \varepsilon)$ rather than $\tilde{O}(1/\sqrt{m})$. The key point is not the specific

estimator, but the existence of an estimator with a clean tradeoff: a finite-sample bound that is explicit in ε and χ, and therefore can be wired into the protocol's completeness and soundness constants.

Why is this enough? The debate protocol only needs the audit to succeed on the *single* step where B chooses to abort. If A is honest, it can arrange that all reported probabilities are approximately correct, and then even a noisy audit will tend to accept. If A is dishonest, B can hunt for a step where A's reported conditional probability is off by a margin. The verifier's audit, if robust, will detect such a deviation with high probability provided the margin exceeds the audit tolerance τ plus the irreducible contamination term $O(\varepsilon)$. Thus, robustness buys us a quantitative statement: we can maintain a constant completeness–soundness gap as long as ε is below a constant threshold and the effective sample size m/χ is large enough. This is the sense in which robust debate is an auditing primitive: it converts an untrusted, strategic computation into a single statistical test whose operating characteristics we can calibrate.

This framing also makes the governance implications clearer. In many proposed oversight deployments, we are not trying to "prove correctness" in the mathematical sense; we are trying to create a credible deterrent against strategic misrepresentation by highly capable models and their operators. A protocol with a constant number of human ratings per task is not merely an asymptotic curiosity: it is the difference between an oversight mechanism that can be integrated into routine model evaluations and one that collapses under scale. At the same time, the formalism reveals the safety tradeoff. If contamination is too high (e.g. a compromised panel) or dependence is too strong (e.g. all raters share the same blind spot), then no single-step audit can reliably distinguish truth from falsehood at constant cost. Robust debate does not eliminate the need for institutional controls on rater quality; it clarifies what those controls must achieve in quantitative terms.

Finally, we should be explicit about what this introduction does *not* claim. We are not assuming that humans can answer arbitrarily complex questions correctly; rather, we assume that the audited query can be presented in a way that elicits a meaningful binary judgment with some stable success probability $p(z)$. We are also not claiming that correlation and contamination are the only pathologies; framing effects, strategic ambiguity, and adversarially chosen queries can all break naive pipelines. Our goal is narrower: to replace a brittle i.i.d. assumption with a deployment-motivated model that captures two dominant failures (dependence and corruption) and to show that, under that model, the debate protocol retains a constant-gap guarantee at constant human cost when parameters are favorable. The next section will recall the baseline protocol structure and highlight exactly where the classical Chernoff argument enters, so that the robust replacement is transparent rather than magical.

2 Background and baseline: stochastic doubly-efficient debate

We now recall the baseline "stochastic doubly-efficient debate" template (as in the source) in the idealized setting where the verifier can query the oracle O directly and obtain independent samples. Our goal in this section is not to re-prove the full theorem, but to isolate (a) the protocol structure that reduces global correctness to a *single* local probability check, (b) the exact point where a Chernoff/Hoeffding argument is invoked, and (c) why a K-Lipschitz stability condition is the right interface between the complexity-theoretic reduction and the statistics of the audit.

2.1 From a probabilistic oracle computation to local conditional probabilities

Fix an input x of length n, and let M be a time-$T(n)$ probabilistic oracle Turing machine deciding L. One can think of an execution of $M^O(x)$ as producing a transcript

$$h_t \;=\; (r_{\le t}, z_{\le t}, o_{\le t})$$

consisting of (some encoding of) internal random coins r, oracle queries z, and oracle replies $o \in \{0,1\}$, revealed over time. The key object for the debate is the Doob martingale of *conditional acceptance probabilities*:

$$p_t \;:=\; \Pr\big[M^O(x) = 1 \,\big|\, h_t\big], \qquad t = 0, 1, \ldots, T,$$

where h_0 is empty and $p_T \in \{0,1\}$ is the realized output bit once the entire transcript is fixed. Intuitively, p_0 is the quantity we ultimately care about (the acceptance probability of M on x), while each increment $p_t - p_{t-1}$ is a *local* update induced by either (i) revealing a fresh random coin or (ii) revealing an oracle answer to some query z_t.

The debate protocol can be read as a mechanism that forces a powerful prover A to expose (approximations to) these p_t's as the computation unfolds, while giving an opposing prover B the ability to "zoom in" on a specific time t where A's asserted local update is inconsistent with the true oracle behavior. The verifier V itself does not try to recompute p_t; it only needs the ability to *audit* one contested local claim.

2.2 Protocol skeleton: sampling a trajectory and allowing a single abort-audit

At a high level, the baseline protocol proceeds along a single sampled trajectory of the computation, maintaining the invariant that if A is honest then the distribution of the transcript matches that of $M^O(x)$. One convenient way to describe the interaction (suppressing some engineering details from

the source, such as discretization and coin-flipping to prevent A from biasing randomness) is:

1. For each round t, A outputs a claimed conditional probability $\hat{p}_t \in [0,1]$ for the next stochastic event relevant to the unfolding transcript (e.g. the probability that the next oracle reply equals 1, conditional on the current history, or the probability that the computation accepts given a branch choice).

2. Using public randomness (or a standard coin-flipping subroutine with A), the verifier realizes an outcome $a_t \in \{0,1\}$ that is *supposed* to be distributed as $\mathrm{Bern}(\hat{p}_t)$. This outcome is appended to the public transcript, and the parties proceed.

3. The challenger B monitors the transcript. At any time t, B may *abort* and point to the current step as the one to audit. The protocol then terminates after a single audit. If B never aborts, the verifier outputs the final realized bit a_T, which (under honesty) is coupled to M's output.

Two features matter for later robustness. First, there is at most *one* empirical contact with the oracle: the single audit triggered by B's abort. Second, the verifier is "doubly-efficient" in the sense that it does not need to simulate M end-to-end; the source protocol includes a cross-examination/localization mechanism ensuring that, when B aborts, the disputed claim can be reduced to an explicit oracle query string z of length l, together with a well-defined target probability $p(z) = \Pr[O(z) = 1]$ that the verifier can estimate by repeated oracle calls. This is where the $\widetilde{O}(l \log T)$ bookkeeping in the verifier running time comes from: the verifier only needs to *index* into the computation and extract the relevant local query/step, not re-run the full computation.

2.3 Where the Chernoff bound enters in the ideal-oracle analysis

In the ideal baseline, the audit is conceptually simple. Suppose B aborts at a step whose correctness reduces to the claim

$$\hat{p}_t \approx p_t,$$

where p_t is, after localization, a conditional probability determined by a single oracle query z (or a constant-size gadget of such queries). The verifier samples the oracle m times on that same z,

$$O(z)^{(1)}, \ldots, O(z)^{(m)} \;\sim\; \text{i.i.d. } \mathrm{Bern}(p(z)),$$

and forms the empirical mean

$$\bar{p} \;:=\; \frac{1}{m}\sum_{i=1}^{m} O(z)^{(i)}.$$

The verifier then compares $\bar{p}$ to $\hat{p}_t$ with a tolerance threshold $\tau > 0$, accepting the disputed step if $|\bar{p} - \hat{p}_t| \le \tau$ and rejecting otherwise.

The only probabilistic estimate needed here is the standard concentration guarantee

$$\Pr\big(|\bar{p} - p(z)| > \tau\big) \;\le\; 2\exp(-2m\tau^2),$$

or an equivalent Chernoff/Hoeffding bound. Choosing

$$m \;=\; \Theta\big(\tau^{-2}\log(1/\delta)\big)$$

ensures that, with probability at least $1-\delta$, the verifier's estimate is τ-accurate. In the baseline soundness argument, this implies: if A reports a value $\hat{p}_t$ that is wrong by more than (roughly) 2τ, then with probability $1-\delta$ the audit detects it and rejects. Conversely, if A is honest, the audit accepts except with probability δ. Thus, the statistical content of the baseline protocol is entirely encapsulated in a one-shot concentration inequality for an empirical mean under i.i.d. sampling.

This is precisely the point of fragility that motivates our later evaluator model: once samples are correlated or partially adversarial, $\bar{p}$ may no longer concentrate at the claimed $\tilde{O}(m^{-1/2})$ rate, and the entire deterrence story ("lying triggers an abort which triggers a reliable audit") can fail.

2.4 Why K-Lipschitzness is the right stability condition

The other key interface is the K-Lipschitz condition on M at the oracle O. Informally, K-Lipschitzness says that small perturbations in the oracle's response probabilities cannot change M's overall acceptance probability by more than K times the size of the perturbation. This is a stability requirement: it rules out computations where an exponentially amplified dependence on a tiny probability discrepancy makes local auditing meaningless.

Operationally, Lipschitzness is what allows the debate to tolerate *approximate* probability checking. The source protocol discretizes claimed probabilities onto a grid of resolution about $1/d$ (with $d = \Theta(K)$), and sets an audit tolerance $\tau = \Theta(1/d)$. The reason is that, if every audited local claim is correct up to $\pm\tau$, then the induced deviation in the global acceptance probability is bounded on the order of $K\tau$, which becomes a constant once we take $\tau = \Theta(1/K)$. In other words, the protocol only needs to distinguish "local error smaller than τ" from "local error larger than τ"; it is not trying to recover probabilities to vanishing precision.

This yields the characteristic quadratic sample scaling in the audit:

$$m \;=\; \Theta\big(\tau^{-2}\log(1/\delta)\big) \;=\; \Theta\big(K^2\log(1/\delta)\big).$$

Crucially, because there is only one audit, this K^2 factor is the *entire* human/oracle sampling cost, independent of the machine's running time T. The complexity-theoretic reduction buys us the fact that we never need to union-bound over T audits; the stability condition buys us the fact that we do not need to set τ to something like $1/T$.

2.5 Completeness and soundness intuition: why one audited step suffices

Finally, we briefly summarize the game-theoretic intuition for why a single audited step is enough to obtain a constant completeness–soundness gap in the ideal model. If A is honest, then (up to discretization) the transcript distribution matches the true stochastic process of $M^O(x)$, so the final bit a_T has essentially the right acceptance probability, and any audit B triggers will pass except with probability δ.

If A is dishonest and aims to shift the verifier's acceptance probability by a constant amount, then (by a hybrid/telescoping argument along the martingale of conditional acceptance probabilities) there must exist some step where A's claimed local conditional probability is off by $\Omega(1/K)$, i.e. by more than the tolerance scale $\tau = \Theta(1/K)$. The role of B is to locate such a step and abort there. Once B aborts, Chernoff concentration ensures the verifier's audit test reliably distinguishes truthful from sufficiently untruthful reporting. Thus the strategic equilibrium is that A is incentivized to stay within the allowed tolerance everywhere, because any significant deviation can be singled out and empirically falsified at bounded cost.

This pinpoints the exact dependency we will exploit next: the baseline proof only needs a guarantee of the form "the audit statistic $\tilde{p}_t$ is within $\pm\tau$ of the true p_t with probability $1-\delta$." In the i.i.d. oracle world, Chernoff provides this immediately for $\tilde{p}_t = \bar{p}$. In the oversight world we care about, we must instead earn an analogous guarantee under contamination and dependence, which is the purpose of the evaluator model introduced in the next section.

3 A clean evaluator model: contamination and dependence

To robustify the audit step, we need a statistical interface that is (i) faithful to how human judgments fail in practice, yet (ii) simple enough to yield finite-sample guarantees that can be plugged into the existing debate reduction. We therefore model the "oracle" not as an algorithm but as a *panel process* that returns multiple noisy binary judgments about the same query. The key move is to separate two qualitatively different deviations from the ideal i.i.d. Bernoulli model: a small *adversarial or arbitrarily biased fraction* of ratings (Huber contamination), and *correlation* among the remaining ratings

(dependence inflation / design effect).

The human-panel oracle $\widetilde{O}$. Fix a query string $z \in \{0,1\}^l$. Under the ideal stochastic oracle O, the verifier would observe a draw in $\{0,1\}$ with success probability

$$p(z) \;:=\; \Pr[O(z) = 1].$$

In the deployment setting we care about, the verifier instead issues z to a panel of m human raters (or, more generally, to m noisy evaluation channels) and receives binary ratings

$$\widetilde{O}(z) \;=\; \big(X_1(z), \ldots, X_m(z)\big) \in \{0,1\}^m.$$

We interpret $X_i(z) = 1$ as "rater i endorses the event whose probability is $p(z)$" (e.g. "the correct oracle reply is 1"). The verifier will aggregate these m bits into an estimate $\tilde{p}(z)$ of $p(z)$ during the single audit. Because the protocol performs at most one audit, it is sufficient—and conceptually clean—to model the distribution of $\widetilde{O}(z)$ *for a fixed audited query* z, without committing to a full joint model across many queries or time steps.

Huber contamination ε: a worst-case fraction of arbitrary ratings. The first failure mode is that some ratings may be uninformative or strategically manipulated. This includes overt adversarial behavior (malicious raters, coordinated brigading, compromised accounts), but also effectively adversarial artifacts (systematic prompt injection into the evaluation interface, labelers misunderstanding the rubric, or a tooling bug that flips outputs). We model this using the standard ε-contamination (Huber) model, per query:

$$X_i(z) \;\sim\; (1-\varepsilon)\, Y_i(z) \;+\; \varepsilon\, Z_i(z),$$

where $Y_i(z) \in \{0,1\}$ is the latent "clean" rating for rater i, and $Z_i(z)$ is an arbitrary $\{0,1\}$-valued draw whose distribution can depend on z and on i. Equivalently, for each i there exists an arbitrary $Q_{z,i} \in \Delta(\{0,1\})$ such that

$$X_i(z) \;\sim\; (1-\varepsilon)\,\mathrm{Bern}\big(p(z)\big) \;+\; \varepsilon\, Q_{z,i}.$$

This is a deliberately pessimistic model: the contaminated samples need not be mean-zero noise, need not be independent, and can be chosen in a way that attempts to pass the audit. In return for this pessimism, we get a clear and interpretable residual term in our estimation guarantees: no estimator can generally identify $p(z)$ to accuracy $o(\varepsilon)$ under ε-contamination, so an $O(\varepsilon)$ additive slack is information-theoretically the right target.

Two practical clarifications are useful. First, ε is not necessarily the literal fraction of "bad people" in the pool; it is a bound on the fraction of *ratings* that behave arbitrarily on the audited question. Second, we allow $Q_{z,i}$ to vary with z, since some tasks are more confusing or more attackable

than others; the protocol and analysis only require a uniform upper bound ε over the class of audited queries.

Dependence inflation χ: effective sample size under correlation. Even when raters are non-adversarial, their judgments are rarely independent. Shared training, shared cultural priors, common failure modes induced by the wording of z, and even real-time coordination can induce positive correlation. If we naively treated $X_1(z), \ldots, X_m(z)$ as i.i.d., we would overstate confidence and potentially make the audit brittle.

We therefore parameterize correlation by a single *design effect* $\chi \geq 1$ that upper-bounds the variance proxy of sums of the clean components. Formally, writing $Y_i(z) \sim \mathrm{Bern}(p(z))$ for the latent uncorrupted ratings, we assume a sub-Gaussian moment bound: for all $\lambda \in \mathbb{R}$ and all subsets $S \subseteq [m]$,

$$\mathbb{E} \exp\Big(\lambda \sum_{i \in S} (Y_i(z) - p(z))\Big) \;\leq\; \exp\Big(\frac{\chi |S| \lambda^2}{2}\Big).$$

When $\chi = 1$, this is consistent with the usual i.i.d. sub-Gaussian behavior of bounded Bernoulli variables. When $\chi > 1$, it asserts that concentration is worse by a factor $\sqrt{\chi}$, which can be read as an *effective sample size* m/χ. This single-parameter summary is exactly what we need downstream: the audit estimator's deviation term will scale like $\sqrt{\chi/m}$ rather than $1/\sqrt{m}$.

We emphasize what this assumption is and is not. It is *not* a claim that ratings follow a Gaussian model, nor that dependence takes a specific form; it is an upper bound on exponential moments of sums, which is the minimal structure required for high-probability bounds. It also cleanly decouples from contamination: χ governs the concentration of the clean part, while ε governs the magnitude of worst-case bias we must tolerate.

Interpretable special cases. The (ε, χ) interface is meant to be operational. Several common panel pathologies can be translated into these parameters.

(1) Independent panel with occasional label noise. If each rater independently produces $Y_i(z) \sim \mathrm{Bern}(p(z))$ and a fraction ε of ratings are arbitrarily corrupted (e.g. random clicks, spam), then $\chi = 1$. This recovers the familiar i.i.d. setting up to an additive $O(\varepsilon)$ bias floor.

(2) Clustered panels / shared latent bias (intra-class correlation). Suppose raters are drawn from clusters (e.g. the same contractor team, the same region, or the same moderation community), and within each cluster judgments are positively correlated because they share a rubric interpretation. A classical approximation is that, for the clean components, the pairwise correlation within a cluster is $\rho > 0$ while different clusters are nearly independent. If the typical cluster size is s, then the variance of the sample

mean inflates by roughly

$$\chi \;\approx\; 1 + (s-1)\rho,$$

the standard design-effect formula from survey sampling. This makes the governance lever explicit: diversifying across clusters (reducing s) or reducing shared bias (reducing ρ) improves χ and therefore reduces the number of ratings m needed for a reliable audit.

(3) Drifting raters / temporally correlated evaluations. In some oversight pipelines, the m "panel" ratings are not collected simultaneously; they are accumulated over time, possibly with rater reuse. Fatigue, learning, or mood can create dependence across a single rater's outputs. Even if each individual rating is marginally unbiased, positive temporal correlation again inflates the variance of the average. Our χ parameter can be read as capturing the worst-case impact of such drift on concentration for the particular audited query, without requiring a full time-series model.

(4) Strategic contamination and partial capture. If an adversary can influence a bounded fraction of ratings (e.g. by inserting compromised evaluators into the pool, or selectively attacking some of the tasks), then ε captures this directly. Crucially, our model allows the contaminated ratings to be *coordinated* and *adaptive* to z, which matches realistic red-team threat models for evaluators.

Why these parameters align with the debate reduction. The debate protocol needs only a one-shot guarantee: when an audit is triggered on some z, the verifier can estimate $p(z)$ to within tolerance τ with failure probability at most δ, and can still detect a prover's misreporting if it exceeds the tolerance by a constant factor. The (ε, χ) model is tailored to yield exactly this kind of bound with the right qualitative tradeoffs:

$$\text{estimation error} \;\sim\; O\Big(\sqrt{\chi/m}\Big) \;+\; O(\varepsilon),$$

so that increasing panel size improves only the stochastic term, while the contamination term imposes a residual bias floor. In particular, because the protocol's tolerance τ is set on the order of $1/K$, we should expect a deployment constraint of the form $\varepsilon \lesssim 1/K$: when the computation is more Lipschitz-sensitive, even a small adversarial bias in the audit can be amplified into a nontrivial completeness–soundness loss.

In the next section we turn this interface into a concrete tool: robust estimators (median-of-means and related variants) that achieve high-probability deviation bounds under ε-contamination and sub-Gaussian dependence inflation χ, with explicit finite-sample scaling suitable for the protocol constants.

4 A robust estimation toolkit for the audit: median-of-means and trimmed means

When the challenger aborts, the verifier must estimate a single Bernoulli mean $p(z)$ from the panel bits $X_1(z), \dots, X_m(z)$ under the (ε, χ) interface. Since the overall debate reduction only ever invokes *one* audit, we can treat the statistical problem as one-shot: given m dependent-and-contaminated samples in $\{0,1\}$, return $\tilde{p}(z)$ such that, with probability at least $1-\delta$,

$$|\tilde{p}(z) - p(z)| \;\le\; \text{(stochastic term)} + \text{(contamination term)} \;\approx\; O\Big(\sqrt{\chi \log(1/\delta)/m}\Big) + O(\varepsilon).$$

This section records two off-the-shelf estimators that achieve exactly this structure with explicit finite-sample constants. We present them in a form tailored to the debate audit: bounded $\{0,1\}$ observations, sub-Gaussian concentration for the clean component with variance proxy χ, and Huber contamination that creates an irreducible $O(\varepsilon)$ bias floor.

Notation for a fixed audited query. Fix the audited query z and write $p := p(z)$, $X_i := X_i(z)$, $Y_i := Y_i(z)$. Thus $Y_i \in \{0,1\}$ has mean p and satisfies the dependence inflation bound summarized by χ, while X_i is the possibly contaminated observation. We suppress z throughout.

4.1 Median-of-means (MoM): a robust estimator with clean constants

The median-of-means estimator uses coarse-graining to isolate the effect of outliers. Intuitively, averaging within blocks controls variance (even under dependence inflation χ), and taking a median across blocks prevents a minority of corrupted blocks from dominating the estimate.

Definition (MoM estimator). Choose an integer $g \ge 2$ (the number of blocks) and let $b := \lfloor m/g \rfloor$ be the block size. Partition $[m]$ into g disjoint blocks $B_1, \dots, B_g$ of size b (ignoring at most $m - gb$ leftover samples). For each block j, compute the block mean

$$\bar{X}_j \;:=\; \frac{1}{b} \sum_{i \in B_j} X_i,$$

and output

$$\tilde{p}_{\mathrm{MoM}} \;:=\; \mathrm{median}\big(\bar{X}_1, \dots, \bar{X}_g\big).$$

Operationally, we can take the blocks to be formed by a fresh random permutation of the m indices before computing block means. (This is not needed for the clean sub-Gaussian concentration, but it is useful in deployments to avoid systematic clustering of correlated raters into the same block.)

Clean concentration within a block. Under the dependence inflation assumption for the clean draws Y_i, each clean block mean concentrates like a sub-Gaussian with variance proxy χ/b. Concretely, for any threshold $t > 0$ and any block B_j,

$$\Pr\big(\,|\bar{Y}_j - p| > t\,\big) \;\le\; 2\exp\Big(-\frac{bt^2}{2\chi}\Big), \qquad \bar{Y}_j := \frac{1}{b}\sum_{i\in B_j} Y_i. \tag{1}$$

This is the precise sense in which χ acts like a variance inflation (or an effective sample size loss): compared to the i.i.d. case, we pay χ multiplicatively in the exponent.

How contamination enters. Because $X_i \in \{0,1\}$, within any block B_j the difference between $\bar{X}_j$ and $\bar{Y}_j$ is controlled by the *fraction of contaminated samples in that block*. Let $C_i \in \{0,1\}$ be the latent indicator that sample i is contaminated (so $X_i = Y_i$ when $C_i = 0$, and X_i is arbitrary when $C_i = 1$). Then

$$|\bar{X}_j - \bar{Y}_j| \;\le\; \frac{1}{b}\sum_{i\in B_j} C_i \;=:\; \alpha_j.$$

In particular, if $\alpha_j \le \alpha$ for many blocks, then those blocks are within α of the corresponding clean block means.

To turn this into a high-probability statement, we use the standard Huber-mixture interpretation in which the contamination indicators are independent Bernoulli(ε) across samples.[1]

Finite-sample MoM bound (explicit constants). Fix a target failure probability $\delta \in (0, 1/2)$. Set

$$g \;:=\; \left\lceil 8\log\frac{2}{\delta}\right\rceil, \qquad b \;:=\; \left\lfloor \frac{m}{g}\right\rfloor, \qquad \tilde{p} \;:=\; \tilde{p}_{\mathrm{MoM}}.$$

Assume $b \ge 32$ and $\varepsilon \le 1/16$. Then

$$\Pr\left(|\tilde{p} - p| > 4\sqrt{\frac{2\chi\log(2/\delta)}{m}} \;+\; 8\varepsilon\right) \;\le\; \delta. \tag{2}$$

The bound has the form we need downstream: a stochastic term scaling as $\sqrt{\chi/m}$ (up to $\sqrt{\log(1/\delta)}$) and an additive contamination term linear in

[1] If one instead allows an adversary to choose an arbitrary subset of εm indices to corrupt after seeing the block partition, then MoM still provides an $O(\varepsilon)$ guarantee (because $\sum_j \alpha_j \le \varepsilon g$), but controlling the *number* of heavily corrupted blocks requires either random blocking hidden from the adversary or an explicit assumption that the contamination indicators are not adversarially coupled to the partition. Since our intended deployment model is a noisy panel rather than an adaptive worst-case scheduler, independent contamination is a clean sufficient condition.

ε. The mild side conditions $b \geq 32$ and $\varepsilon \leq 1/16$ are consistent with our intended regime: b is a constant once we choose $m = \Theta(\chi \log(1/\delta))$, and ε must be a small constant anyway to preserve a constant completeness–soundness gap.

Proof sketch (what is doing the work). There are two ingredients. First, by (1) with $t := 2\sqrt{2\chi \log(2/\delta)/m}$ (which is on the order of $\sqrt{\chi \log(1/\delta)/(gb)}$), most clean block means $\bar{Y}_j$ lie within t of p. Second, by a Chernoff bound on $\alpha_j = b^{-1} \sum_{i \in B_j} C_i$ with $C_i \sim \mathrm{Bern}(\varepsilon)$, most blocks satisfy $\alpha_j \leq 2\varepsilon$. On the intersection of these two majority events, at least half the observed block means $\bar{X}_j$ lie within $t + 2\varepsilon$ of p, hence their median $\tilde{p}$ lies within the same interval (up to constant slack absorbed into (2)).

4.2 Trimmed means: simpler aggregation and similar guarantees

Median-of-means is robust but can be slightly discontinuous as a function of the data. A closely related alternative is to compute block means and then take a *trimmed mean* across blocks, which can be easier to implement and tune in practice (e.g. when we want a smooth dependence on samples for monitoring).

Definition (blockwise trimmed mean). As above, partition into g blocks and compute $\bar{X}_1, \ldots, \bar{X}_g$. Fix a trimming fraction $\alpha \in (0, 1/2)$. Let $\bar{X}_{(1)} \leq \cdots \leq \bar{X}_{(g)}$ denote the sorted block means, and define

$$\tilde{p}_{\mathrm{trim}} \;:=\; \frac{1}{g - 2\lfloor \alpha g \rfloor} \sum_{j=\lfloor \alpha g \rfloor + 1}^{g - \lfloor \alpha g \rfloor} \bar{X}_{(j)}.$$

We will use $\alpha = 1/4$, which trims away the most suspicious quarter of blocks on each side.

Finite-sample trimmed-mean bound (explicit constants). With g, b as above and trimming $\alpha = 1/4$, under the same side conditions $b \geq 32$ and $\varepsilon \leq 1/16$, we have

$$\Pr\left(|\tilde{p}_{\mathrm{trim}} - p| > 6\sqrt{\frac{2\chi \log(2/\delta)}{m}} \;+\; 12\varepsilon \right) \;\leq\; \delta. \tag{3}$$

The constants are slightly worse than MoM (reflecting the fact that trimming averages the surviving blocks, so a few moderately biased blocks can still shift the estimate), but the same qualitative tradeoff holds, and in practice the trimmed mean is often numerically stable.

4.3 A plug-in corollary for the audit threshold

The debate audit does not require vanishing estimation error; it requires that, for a tolerance τ (later set to $\Theta(1/K)$), the verifier can enforce

$$|\tilde{p} - p| \;\le\; \tau \quad \text{with probability at least } 1 - \delta,$$

up to an unavoidable $O(\varepsilon)$ slack. Combining (2) (or (3)) with a simple rearrangement yields the parameterization we will use in the protocol.

Corollary (sample size sufficient for τ-accuracy). Using median-of-means, if

$$m \;\ge\; 128\,\chi\,\tau^{-2}\,\log\frac{2}{\delta} \quad \text{and} \quad \varepsilon \;\le\; \frac{\tau}{16},$$

then the audit estimate $\tilde{p}$ satisfies

$$\Pr\big(|\tilde{p} - p| > \tau\big) \;\le\; \delta.$$

Thus, once we fix the protocol's tolerance τ, the required number of human ratings scales as $\Theta(\chi\tau^{-2}\log(1/\delta))$, matching the "effective sample size" intuition and making explicit how correlation (χ) and desired confidence (δ) translate into audit cost. In the next section we instantiate τ in terms of the machine's Lipschitz sensitivity K, and we show how this one-shot estimate suffices to robustify the entire abort-audit step.

5 Robustified debate protocol: modifying the abort–audit step

We now describe the protocol-level modification that turns the one-shot robust mean estimators from the previous section into a verifier that tolerates (ε, χ)-imperfect human panels. Conceptually, nothing about the stochastic doubly-efficient debate changes *except* what happens when the challenger aborts: instead of treating the panel as an i.i.d. Bernoulli oracle and applying a Chernoff-style check, we (i) aggregate the m binary ratings with a robust estimator (median-of-means or trimmed mean), and (ii) choose the tolerance τ and sample size m so that this robust estimate is accurate enough at the *single* audited step.

5.1 What the verifier needs from the audit

Recall the role of the audit in the underlying debate reduction. At each debate round t, prover A announces a conditional probability $\hat{p}_t \in [0, 1]$ for the next oracle bit in the stochastic simulation (formally, a conditional probability of the form $p_t = \Pr[O(z_t) = 1 \mid \text{transcript}_{<t}]$ for an implicit query string z_t determined by the transcript). Prover B monitors the transcript

and can *abort* at some round t to force an audit of that specific conditional probability claim.

At an abort round t, the verifier must implement the following predicate with high probability: accept if $|\hat{p}_t - p_t|$ is small, and reject if $|\hat{p}_t - p_t|$ is noticeably large. Importantly, we do *not* need to estimate p_t to vanishing error; we only need separation at a tolerance scale τ that is later tied to the discretization in the debate protocol. The previous section gives exactly the right tool: for a single query z, from m bits $X_1(z), \ldots, X_m(z)$ we can compute $\tilde{p}(z)$ such that, with probability at least $1 - \delta$,

$$|\tilde{p}(z) - p(z)| \;\le\; O\Big(\sqrt{\chi \log(1/\delta)/m}\Big) + O(\varepsilon).$$

The additive $O(\varepsilon)$ term is information-theoretically unavoidable under Huber contamination, so we will explicitly budget for it in the audit threshold.

5.2 Protocol modification: robust audit rule

Fix an input instance x of length n. The robustified debate protocol is identical to the source stochastic doubly-efficient debate up to the abort rule, with the following replacement.

Robust audit subroutine at round t. Upon abort by B at round t, the verifier:

1. Identifies the audited query string $z_t \in \{0,1\}^l$ implicit in the transcript (the same query that the source protocol would audit at this point).

2. Queries the human-panel oracle $\widetilde{O}$ on z_t to obtain m binary ratings $X_1(z_t), \ldots, X_m(z_t)$.

3. Computes a robust estimate $\tilde{p}_t := \tilde{p}(z_t)$ using either:

   - *Median-of-means:* partition the m indices into g blocks, average within blocks, and take the median; or
   - *Blockwise trimmed mean:* partition into g blocks, average, sort the block means, and average after trimming a constant fraction.

   We recommend median-of-means for analysis and trimmed means for monitoring; either satisfies the required deviation inequality with the same scaling.

4. Accept the audit if
$$|\tilde{p}_t - \hat{p}_t| \;\le\; \tau, \tag{4}$$
and reject otherwise.

If no abort occurs, the verifier follows the source protocol and outputs the final sampled bit a_T.

This rule is deliberately minimal: it preserves the *single-audit* structure that makes the debate doubly efficient, and it localizes all human imperfection into a one-shot statistical guarantee.

5.3 Setting the tolerance τ via discretization and Lipschitzness

The remaining question is how to set τ. In the source construction, probabilities are discretized to a grid of width $\Theta(1/d)$, where d is chosen proportional to the Lipschitz constant K of the oracle machine M. Intuitively, K-Lipschitzness says that if we perturb the oracle answers by a small amount at each query, the overall acceptance probability of $M(x)$ changes by at most K times the perturbation magnitude. Thus, if we enforce that each audited conditional probability is correct to within $\Theta(1/K)$, then the induced perturbation on the final acceptance probability is $O(1)$, which is exactly what we need to maintain a constant completeness–soundness gap.

Concretely, we set

$$d \;:=\; \lceil c_d K \rceil\,, \qquad \tau \;:=\; \frac{c_\tau}{d},$$

for absolute constants $c_d, c_\tau > 0$ chosen to match the source protocol's discretization and to leave slack for the robust estimation error. Since $d = \Theta(K)$, this enforces $\tau = \Theta(1/K)$. The only additional requirement induced by human imperfection is that the contamination floor $O(\varepsilon)$ must be below the tolerance scale: we will assume

$$\varepsilon \;\le\; c_\varepsilon \tau \;=\; \Theta(1/K), \tag{5}$$

for a sufficiently small absolute constant $c_\varepsilon > 0$. This is the operational form of the condition "contamination must be small enough to be auditable."

5.4 Choosing the panel size m and block count g

Given τ and a target audit failure probability $\delta \in (0, 1/2)$, we choose m so that the robust estimate $\tilde{p}_t$ is τ-accurate up to the contamination floor. Using the plug-in corollary from the previous section (stated there with explicit constants for median-of-means), it suffices to take

$$m \;\ge\; c_m\, \chi\, \tau^{-2}\, \log \frac{2}{\delta}, \tag{6}$$

for a large enough absolute constant c_m, together with the side condition $\varepsilon \le c_\varepsilon \tau$ as in (5). Substituting $\tau = \Theta(1/K)$ yields the headline scaling

$$m \;=\; \Theta\Big(\chi K^2 \log(1/\delta)\Big),$$

which is the only place where human cost enters the debate.

For median-of-means (or trimmed means on block averages), we also set the number of blocks

$$g \;:=\; \left\lceil c_g \log \frac{2}{\delta} \right\rceil, \qquad b \;:=\; \left\lfloor \frac{m}{g} \right\rfloor,$$

with c_g a sufficiently large absolute constant so that a median argument drives the failure probability down to δ. In implementation, we can randomly permute rater indices before blocking; this does not change correctness under our χ-sub-Gaussian assumption, but it reduces the chance that a correlated subgroup is concentrated into a single block in deployments.

5.5 Why this is the only change we need

The core debate analysis (carried out in the next section) reduces completeness and soundness to the existence of a single round t where B can force an informative audit if A deviates from truthful conditional probabilities. Our robust audit rule preserves exactly that structure. The key property we will repeatedly invoke is the following implication, which is immediate from the deviation inequality for $\tilde{p}_t$ and a triangle inequality:

Audit separation guarantee (informal). If at an abort round t the true conditional probability is p_t and A misreports by more than the estimator's error budget—namely if

$$|\hat{p}_t - p_t| \;\geq\; 2\tau + O(\varepsilon),$$

then with probability at least $1 - \delta$ the verifier rejects the audit because $|\tilde{p}_t - \hat{p}_t| > \tau$. Conversely, if A is truthful up to τ and ε-slack, the verifier accepts with probability at least $1 - \delta$.

This is precisely the "local soundness" condition that the source protocol needs at the abort point, except that (i) we pay an additive $O(\varepsilon)$ degradation, and (ii) we replace i.i.d. Chernoff tails by χ-inflated concentration and a robust estimator. Since the debate triggers at most one audit, we do not need to union bound over many human queries; δ can therefore be treated as a protocol constant, and the total number of human ratings is $q = m$ up to constant factors.

5.6 Complexity and deployment interpretation

The verifier's running time increases only by the cost of computing $\tilde{p}_t$ from m bits: median-of-means and trimmed means run in time $\widetilde{O}(m)$ (dominated by sorting $g = \Theta(\log(1/\delta))$ block means if we implement trimming, or by selecting the median). Thus the overall verifier time becomes

$$\widetilde{O}(l \log T + m) \;=\; \widetilde{O}\Big(l \log T + \chi K^2 \log(1/\delta)\Big),$$

and, critically, the human cost m is independent of the computation length T.

From an oversight perspective, χ summarizes correlation and shared-mode failure in the panel (a design effect), while ε summarizes the irreducible adversarial or systematically biased fraction. The protocol's parameterization makes the tradeoff explicit: if we can diversify panels to reduce χ, we reduce required m linearly; but if ε is too large relative to $\tau = \Theta(1/K)$, then no choice of m can recover a clean separation at the required tolerance scale, and the debate's constant gap necessarily collapses. The next section formalizes this in the completeness–soundness theorem by propagating the τ-level audit guarantee through the stochastic debate induction.

6 Main theorem: completeness and soundness under (ε, χ)-imperfect panels

We now state and prove the quantitative guarantee that the robustified abort–audit rule suffices to recover a constant completeness–soundness gap while keeping the total number of human ratings independent of the simulated runtime T. The high-level structure of the argument is unchanged from the source stochastic doubly-efficient debate: we still reduce global correctness of the debate outcome to the existence of a *single* round t at which (i) any meaningful misreport by A induces a local discrepancy in a conditional probability claim, and (ii) B can force an audit at such a round. The only new work is to show that our robust estimator-based audit instantiates the same local predicate, up to an additive degradation proportional to ε and δ.

Theorem (robust stochastic doubly-efficient debate). Fix an input $x \in \{0,1\}^n$, and let M be a time-$T(n)$ probabilistic oracle Turing machine with query length l, which is K-Lipschitz at an ideal stochastic oracle O. Suppose the human-panel oracle $\widetilde{O}$ satisfies the Huber contamination model with rate ε and the χ-inflated sub-Gaussian dependence condition as specified in the enclosing scope. Consider the robustified protocol in which, upon abort at time t, the verifier estimates $p_t = \Pr[O(z_t) = 1 \mid \text{transcript}_{<t}]$ by a robust mean estimator $\tilde{p}_t$ from m panel ratings and accepts iff $|\tilde{p}_t - \hat{p}_t| \le \tau$, where $\tau = \Theta(1/K)$.

There exist absolute constants $c_0, C_1, C_2 > 0$ such that if $\varepsilon \le c_0/K$ and

$$m \;\ge\; c_m\, \chi\, \tau^{-2}\, \log\frac{2}{\delta} \;=\; \Theta\big(\chi K^2 \log(1/\delta)\big),$$

then the protocol decides L with

$$x \in L \implies \Pr[V=1] \ge \frac{3}{5} - C_1\varepsilon - C_2\delta, \qquad x \notin L \implies \Pr[V=1] \le \frac{2}{5} + C_1\varepsilon + C_2\delta,$$

and the verifier runs in time $\widetilde{O}(l \log T + m)$ while using at most $q = m$ total human ratings.

6.1 Proof strategy: reduce to a single audited predicate

The source debate analysis can be viewed as an induction over rounds $t = 1, \ldots, T$ that compares two processes: (i) a "truthful" stochastic simulation driven by the ideal oracle probabilities p_t, and (ii) the transcript distribution induced when A announces $\hat{p}_t$ and the protocol samples the next bit a_t using $\hat{p}_t$ (together with the standard cross-examination steps that ensure $\hat{p}_t$ is binding). The key technical invariant is that if all conditional probabilities are accurate on a discretization grid at tolerance τ, then the final acceptance probability differs by at most $O(K\tau)$, which we choose to be a sufficiently small absolute constant by setting $\tau = \Theta(1/K)$.

Thus, to preserve the source completeness–soundness gap, it suffices that the abort–audit step implements a *local* correctness test: whenever $|\hat{p}_t - p_t|$ is noticeably larger than τ, an abort at round t causes rejection with probability at least $1 - \delta$; and whenever $|\hat{p}_t - p_t|$ is at most τ (up to the unavoidable ε-bias), the audit accepts with probability at least $1 - \delta$. This is exactly the role played by Chernoff bounds in the i.i.d. oracle model, and it is exactly what robust mean estimation restores under (ε, χ)-imperfect panels.

6.2 Step 1: local audit correctness with contamination and dependence

Fix a round t at which B aborts, and write z_t for the audited query determined by the transcript. Let $p_t := p(z_t)$ denote the ideal conditional probability of the audited bit under O. The verifier forms $\tilde{p}_t$ from m ratings $X_1(z_t), \ldots, X_m(z_t)$ using median-of-means (or an equivalent robust blockwise estimator). By the robust mean estimation lemma stated earlier, for a suitable absolute constant $C > 0$,

$$\Pr\Big(|\tilde{p}_t - p_t| > C\Big(\sqrt{\chi \log(2/\delta)/m} + \varepsilon\Big)\Big) \le \delta. \tag{7}$$

Choosing m as in the theorem makes the stochastic term at most τ (up to constant factors), and our standing assumption $\varepsilon \le c_0/K = \Theta(\tau)$ ensures the contamination floor is also within the audit slack budget. Concretely, we can pick constants so that

$$C\Big(\sqrt{\chi \log(2/\delta)/m} + \varepsilon\Big) \;\le\; \tau.$$

On this event, we have a clean triangle-inequality implication:

$$|\hat{p}_t - p_t| \ge 2\tau \quad \Longrightarrow \quad |\tilde{p}_t - \hat{p}_t| \ge \tau, \tag{8}$$

so the verifier rejects the audit with probability at least $1 - \delta$. Conversely, if $|\hat{p}_t - p_t| \le \tau$, then $|\tilde{p}_t - \hat{p}_t| \le 2\tau$ on the same good event, and by slightly widening constants in the accept region (as the source protocol already does via discretization slack), we obtain acceptance with probability at least $1-\delta$. The only substantive difference from the i.i.d. setting is that the "good event" now costs δ and its radius contains an irreducible $O(\varepsilon)$ term.

6.3 Step 2: plug the local predicate into the source induction

We now treat the robust audit guarantee as a black-box replacement for the source "audit succeeds with overwhelming probability" claim. Since the protocol performs *at most one* audit, we do not pay a union bound over t; δ enters only once, additively, and can be interpreted as a verifier-level constant rather than something that scales with T.

Completeness follows by exhibiting a polynomial-time A strategy $A^\star$ for $x \in L$: namely, $A^\star$ announces the discretized truthful probabilities $\hat{p}_t$ (rounded to the protocol grid of width $\Theta(1/d) = \Theta(\tau)$) and otherwise follows the source protocol. In the no-abort branch, the distribution of a_T is within $O(K\tau)$ total variation of the ideal simulation, hence yields acceptance probability at least $3/5$ after choosing constants in $d = \Theta(K)$. If B aborts, then by construction $|\hat{p}_t - p_t| \le O(\tau)$ for every t, so the robust audit accepts with probability at least $1 - \delta$, again up to discretization slack. The only new degradation is that the truthful discretization must be wide enough to absorb the unavoidable $O(\varepsilon)$ bias from (7); this is where the additive $-C_1\varepsilon$ term enters, and why we require $\varepsilon \le \Theta(\tau) = \Theta(1/K)$ to prevent the local test from becoming information-theoretically impossible.

Soundness is symmetric: for $x \notin L$, the source analysis constructs a polynomial-time challenger $B^\star$ that monitors the transcript and aborts at the first round t where A's implied conditional probability claim must deviate from the truth by $\Omega(\tau)$ in order to maintain the appearance of a globally accepting computation. The existence of such a round is exactly what the source induction proves, using the discretization and the K-Lipschitz bound to argue that if all $|\hat{p}_t - p_t|$ were small then the final acceptance probability could not cross the constant gap. Once $B^\star$ aborts at such a round, (8) implies the robust audit rejects with probability at least $1-\delta$, except for the additive slack due to ε. This yields $\Pr[V = 1] \le 2/5 + C_1\varepsilon + C_2\delta$ after accounting for the single audit-failure event and the same $O(K\tau)$ discretization error.

6.4 Verifier complexity and the "constant-cost" interpretation

The verifier's additional work beyond the source protocol is confined to the one-shot computation of $\tilde{p}_t$ from m bits, which can be done in $\widetilde{O}(m)$ time (partitioning plus median selection, or sorting $g = \Theta(\log(1/\delta))$ block means).

Hence total verifier time is $\widetilde{O}(l \log T + m)$, and total human ratings are exactly $q = m$, independent of T. From a deployment perspective, this is the point at which the formalism reveals the safety tradeoff: we can buy down correlation in the panel by lowering χ (e.g. by diversifying rater pools), thereby reducing cost linearly, but we cannot buy down ε with more samples beyond the additive floor. The protocol therefore cleanly separates two oversight failure modes—shared-mode dependence (χ) and adversarial/systems-level bias (ε)—and makes explicit when each becomes the binding constraint.

Finally, we emphasize a limitation that will matter in governance settings: the theorem guarantees a constant gap only under $\varepsilon \leq \Theta(1/K)$. When K is large (high sensitivity of M to oracle perturbations), the required audit tolerance τ is correspondingly small, and even modest systematic bias in human judgments can swamp the audit signal. The next section formalizes this intuition as a tightness result: beyond constant thresholds in ε (or when χ effectively grows with task difficulty), any single-audit constant-cost scheme must fail.

7 Tightness and limits: when constant-cost single-audit oversight must fail

The previous section makes a strong claim: a verifier can certify an arbitrarily long stochastic computation while paying for only a single, constant-sized human audit. This is only plausible because the protocol reduces global correctness to a *local* probability check with tolerance $\tau = \Theta(1/K)$, and because median-of-means turns m noisy panel bits into an estimate $\tilde{p}_t$ that is accurate up to $\sqrt{\chi/m}$ plus an irreducible ε-bias. In this section we justify that these are not merely artifacts of analysis: if the contamination rate ε is too large, or if the dependence inflation χ grows so that the effective sample size is bounded, then *any* protocol that performs only one such audit with $O(1)$ total ratings cannot maintain a constant completeness–soundness gap uniformly over all tasks.

7.1 Why ε creates an information-theoretic floor

At a high level, single-audit debate reduces to the following statistical primitive. Conditioned on the transcript up to the abort time t, there is a fixed query z_t and a fixed (ideal) Bernoulli parameter $p_t = p(z_t)$. The verifier observes m binary ratings $X_1(z_t), \ldots, X_m(z_t)$ drawn from an (ε, χ)-imperfect panel, and must decide whether A's reported $\hat{p}_t$ is within tolerance τ of p_t. No matter what estimator or test we use, we cannot beat the minimax distinguishability limits of contaminated Bernoulli samples.

A simple way to see the floor is to consider worst-case Huber contamination with no dependence (set $\chi = 1$ for this argument). Fix two Bernoulli

parameters p and p' with $|p - p'| \le 2\varepsilon$. Define contamination distributions on $\{0, 1\}$ by choosing

$$Q(1) = \frac{p' - (1-\varepsilon)p}{\varepsilon}, \qquad Q'(1) = \frac{p - (1-\varepsilon)p'}{\varepsilon},$$

which lie in $[0, 1]$ precisely when $|p - p'| \le 2\varepsilon$. Then the two contaminated marginals coincide:

$$(1-\varepsilon)\,\mathrm{Bern}(p) + \varepsilon\, Q \;=\; (1-\varepsilon)\,\mathrm{Bern}(p') + \varepsilon\, Q'.$$

Consequently, *for every* sample size m and *every* estimator $\widehat{p} = \widehat{p}(X_1, \ldots, X_m)$, the joint distributions of $(X_1, \ldots, X_m)$ under the two hypotheses are identical (in fact product-identical), and no procedure can reliably tell whether the underlying ideal mean was p or p'. In particular, for any target accuracy $\alpha < \varepsilon$, there exist p, p' separated by 2α such that

$$\inf_{\widehat{p}} \sup_{\text{Huber }(\varepsilon)} \Pr\left(|\widehat{p} - p| > \alpha\right) \;\ge\; \frac{1}{2}.$$

This demonstrates the qualitative tightness of the additive ε term in robust mean estimation: beyond constant factors, contamination behaves like an *unremovable* shift in the parameter.

Two corollaries matter for our oversight setting. First, if $\varepsilon \ge 1/2$, then by the same construction (taking $p = 0$, $p' = 1$) the contaminated marginal can be made identical even when the ideal oracle answers are maximally different; therefore no single-audit protocol with a bounded number of ratings can maintain any nontrivial completeness–soundness gap across all inputs. Second, even when $\varepsilon < 1/2$, any local predicate that attempts to certify correctness to tolerance τ must have $\tau \gtrsim \varepsilon$ in the worst case. Since our global reduction forces $\tau = \Theta(1/K)$, we obtain a necessary condition of the form

$$\varepsilon \;\le\; c\,\tau \;=\; \Theta\!\left(\frac{1}{K}\right),$$

matching the theorem's standing assumption up to constants. Intuitively: when the simulated computation is highly sensitive (large K), the protocol demands extremely fine-grained probability checking, and adversarial or systematically biased human judgments can swamp that signal even if we query more people.

7.2 Dependence inflation and the collapse of effective sample size

The second failure mode is not bias but *correlation*. The dependence parameter χ enters our concentration bounds exactly as a variance inflation (a design effect): the stochastic term in (7) behaves like $\sqrt{\chi/m}$ rather than

$1/\sqrt{m}$. This scaling is also tight in the sense that, without further assumptions, χ can render additional panelists almost useless.

To make this concrete, consider an exchangeable "copying" model at a fixed audited query z: with probability ρ, all m raters output the *same* latent draw $Y \sim \mathrm{Bern}(p)$ (perfect within-panel correlation), and with probability $1-\rho$ they are i.i.d. $\mathrm{Bern}(p)$. Then the variance of the sample mean $\bar{X}$ is

$$\mathrm{Var}(\bar{X}) = p(1-p)\left(\frac{1-\rho}{m} + \rho\right),$$

so once ρ is bounded away from 0, $\mathrm{Var}(\bar{X}) = \Theta(1)$ regardless of m. In the language of design effects, χ is on the order of $1 + (m-1)\rho$, so achieving $\sqrt{\chi/m} \ll 1$ requires $\rho \ll 1/m$. Said differently, if the panel is effectively one shared latent opinion (because of common framing, common sources, or outright coordination), then we do not have m independent bits of evidence; we have $O(1)$.

This yields an impossibility statement for constant-cost single-audit oversight: suppose we wish to distinguish $p = 1/2 + \gamma$ from $p = 1/2 - \gamma$ at constant advantage. Under standard two-point testing lower bounds (e.g. Le Cam's method), the required number of *effective* independent samples is $\Omega(1/\gamma^2)$. Under dependence inflation, m panel ratings supply only m/χ effective samples, so any audit that succeeds with constant probability must satisfy

$$m \;\geq\; \Omega\!\left(\frac{\chi}{\gamma^2}\right).$$

In our protocol the relevant separation scale is $\gamma = \Theta(\tau)$, hence $m = \Omega(\chi\,\tau^{-2}) = \Omega(\chi K^2)$ is not merely sufficient but essentially necessary. Therefore, if χ is allowed to grow with task difficulty in such a way that χK^2 is unbounded, then the "constant-cost" promise breaks: either the verifier must increase m accordingly, or accept that the audit frequently cannot tell truth from misreport.

7.3 Why χ can scale with T in realistic deployments

The theoretical role of χ is to summarize whatever dependence structure remains after we choose a panel and a presentation format for z_t. In deployments, there are at least three reasons χ may effectively scale with the underlying runtime T, even though the audit targets only one step.

First, query *complexity* can induce shared heuristics: as the audited step becomes more abstract or underspecified, raters may converge on the same superficial cue, increasing intra-class correlation. Second, *selection* effects can couple rater errors across tasks: if the same pool of experts repeatedly audits related queries, their shared background (and shared blind spots) induces positive correlation that is hard to average out. Third, *strategic*

dependence can arise under adversarial pressure: if there exist channels by which raters can be influenced (common media narratives, coordinated campaigns, or organizational capture), then many ratings can shift together, again behaving like a large χ even without explicit collusion.

These observations connect the mathematical limit to a governance implication. The debate protocol by itself does not create independence; it merely exploits it. If we cannot design oversight so that different ratings are meaningfully de-correlated (small χ) and not systematically biased (small ε), then there is no purely algorithmic fix that preserves constant human cost while maintaining a constant correctness gap.

7.4 Takeaway for the next section

The theorem and the limits together give a crisp agenda for calibration. The only way the single-audit idea remains economically attractive is if, in the targeted domain, we can empirically bound (or actively control) both (i) the contamination-like bias floor ε relative to the required tolerance $\tau = \Theta(1/K)$, and (ii) the design effect χ so that $m = \Theta(\chi K^2 \log(1/\delta))$ is practically small. The next section therefore turns from worst-case theory to measurement: we specify clustered-rater generative models, estimate intra-class correlation and map it to χ, and numerically explore how audit size and error change as we vary diversification, instruction design, and rater-pool structure.

8 Practical calibration and simulations (numerical): clustered-rater models, mapping ICC to χ, and audit sensitivity

The previous section reduces "single-audit oversight" to two measurable quantities: a bias floor (captured by ε) and an effective sample-size collapse factor (captured by χ). What we do *not* get for free from the theory is a trustworthy value of χ (or ε) for a given deployment. In practice, both parameters are properties of a *panel design* (who we recruit, what we show them, how we randomize prompts, and how we aggregate), not just properties of the underlying task. This section therefore treats calibration explicitly as an empirical–numerical step: we specify simple clustered-rater generative models that expose the relevant failure modes, show how to estimate intra-class correlation (ICC) and translate it into a conservative χ, and run sensitivity analyses that convert measured $(\widehat{\varepsilon}, \widehat{\chi})$ into audit sizes m and expected error.

8.1 Clustered-rater generative models for a single audited query

Fix an audited query z with ideal oracle parameter $p = p(z)$. We want a model family that separates (i) *shared* sources of variance (framing effects, common heuristics, common misinformation) from (ii) *idiosyncratic* rater noise, and then overlays (iii) adversarial or arbitrarily biased contamination.

A convenient abstraction is a two-level random-effects model on the latent "uncorrupted" ratings:

$$\theta \sim \mathcal{P}_p, \qquad Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta), \qquad X_i \sim (1-\varepsilon)Y_i + \varepsilon Z_i,$$

where $\mathcal{P}_p$ is a distribution on $[0,1]$ with mean p (capturing a shared latent "panel state"), and Z_i is arbitrary (capturing Huber contamination).

Two concrete special cases are useful in simulations.

Copying / common-shock model (exchangeable, interpretable ICC). Let $C \sim \text{Bern}(\rho)$ and $Y \sim \text{Bern}(p)$. If $C = 1$, set $Y_i = Y$ for all i (a perfect common shock); if $C = 0$, draw $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. This produces exact exchangeability and an ICC equal to ρ for the latent Y_i (up to the usual $p(1-p)$ scaling). It captures the extreme regime where correlation is caused by "everyone using the same cue."

Beta–binomial model (continuous shared uncertainty). Let $\theta \sim \text{Beta}(\alpha, \beta)$ with $\mathbb{E}[\theta] = p$. Then $Y_i \mid \theta$ are i.i.d. Bernoulli. This corresponds to a softer shared effect: the panel collectively drifts toward optimism or pessimism about the query, then individuals add independent noise. Writing $\mu = p$ and $\sigma_\theta^2 = \text{Var}(\theta)$, one obtains

$$\text{Cov}(Y_i, Y_j) = \sigma_\theta^2, \qquad \text{Var}(Y_i) = \mu(1-\mu),$$

and therefore an ICC (for the latent Bernoulli draws) of

$$\rho \;=\; \frac{\text{Cov}(Y_i, Y_j)}{\text{Var}(Y_i)} \;=\; \frac{\sigma_\theta^2}{\mu(1-\mu)}.$$

This ρ increases when shared framing dominates individual-level noise.

We emphasize that these are *measurement* models: they are not claims about human cognition, but convenient ways to parameterize how quickly additional raters stop helping.

8.2 From ICC to a conservative χ

The analysis in the previous sections packages dependence into a design-effect-like χ that inflates concentration as if we had only m/χ effective samples. In the common exchangeable case (a good first approximation when

raters are drawn from a homogeneous pool under a shared prompt), χ can be tied directly to ICC.

If ρ denotes the ICC of the (uncorrupted) Bernoulli ratings, then the variance of the sample mean satisfies the familiar identity

$$\mathrm{Var}\big(\bar{Y}\big) = \frac{\mu(1-\mu)}{m}\big(1 + (m-1)\rho\big).$$

Comparing to the independent case $\mathrm{Var}(\bar{Y}) = \mu(1-\mu)/m$ suggests the mapping

$$\chi(m) \;\approx\; 1 + (m-1)\rho.$$

This is not merely algebraic convenience: it makes the core deployment lesson explicit. When ρ is bounded away from 0, $\chi(m)$ grows linearly with m, so $\sqrt{\chi(m)/m} = \sqrt{\rho + (1-\rho)/m}$ plateaus at $\sqrt{\rho}$. No estimator can average away a *shared* error mode.

For protocol parameter-setting, we recommend using a *conservative* χ derived from an upper confidence bound on ρ, evaluated at the intended audit size m. If we target a tolerance τ, then the correlation plateau yields an immediate feasibility check:

$$\sqrt{\rho} \;\lesssim\; \tau \quad\Longleftrightarrow\quad \rho \;\lesssim\; \tau^2.$$

If empirical calibration finds $\rho > \tau^2$ for the current prompt and pool, then increasing m cannot achieve the desired local accuracy, and the only remedies are to (i) redesign the panel to reduce ρ (diversify, change framing, isolate raters), or (ii) relax τ (which in our setting is constrained by K), or (iii) abandon single-audit constant-cost ambitions for that domain.

8.3 Estimating ICC (and thus χ) from pilot audits

A practical calibration protocol is:

1. **Collect pilot queries.** Sample a set of representative audited-query templates z from the intended application distribution (including "hard" ambiguous cases, since they often drive correlation).

2. **Run repeated panels.** For each z, solicit ratings from multiple raters across multiple pools (e.g. contractors, experts, geographically separated groups), recording pool identity and minimal metadata needed for clustering.

3. **Fit a clustered binary model.** For binary ratings, a mixed-effects logistic model is often more stable than ANOVA-style ICC:

$$\mathrm{logit}\ \Pr[X_{z,i} = 1] = \beta_z + u_{\mathrm{pool}(i)} + v_i,$$

where u is a pool-level random effect (shared), and v_i is an individual random effect. One can then compute an implied ICC on the latent-logit scale and translate it to an approximate probability-scale ρ by simulation (parametric bootstrap).

4. **Convert to χ.** Use $\widehat{\rho}$ (or an upper confidence bound) to set $\widehat{\chi}(m) = 1+(m-1)\widehat{\rho}$ for the planned m, or more conservatively $\widehat{\chi} = \sup_{m \le m_{\max}}(1+(m-1)\widehat{\rho})$ if m may vary.

Two cautions matter. First, ICC is often heterogeneous across query types; using a single global χ is safe only if it is pegged to a high quantile of the observed ICC distribution. Second, contamination can inflate apparent ICC if adversarial raters coordinate; this is not a nuisance but a signal that the procurement process is failing in a strategically relevant way.

8.4 Numerical sensitivity: translating (ε, χ) into audit sizes and error

Given $(\widehat{\varepsilon}, \widehat{\chi})$, the theorem-level guidance for audit sizing is

$$m \;\approx\; c\,\widehat{\chi}\,\tau^{-2}\log(2/\delta),$$

up to constants depending on the robust estimator and the precise dependence model. Numerically, we care about three linked outputs: (i) the empirical deviation $\Pr(|\tilde{p} - p| > \tau)$ under the panel model, (ii) the audit test power $\Pr(|\tilde{p} - \hat{p}| \ge \tau)$ when $\hat{p}$ is wrong by a margin, and (iii) the realized completeness–soundness gap once these local events are plugged into the global protocol.

A simple simulation harness for a fixed query proceeds as follows. Choose $p \in [0,1]$ (typically $p = 1/2$ is worst-case for variance), choose a dependence parameter (e.g. ρ in the copying model or (α, β) in the beta–binomial), choose ε and a contamination strategy (e.g. flipping to maximize bias relative to p), draw m ratings, compute $\tilde{p}$ using (a) the sample mean and (b) median-of-means with $g \asymp \log(2/\delta)$ blocks, and record whether the audit would accept for a given $\hat{p}$.

Across such simulations, three robust qualitative patterns recur.

(1) Robust aggregation mainly buys protection against ε, not against ρ. Median-of-means (and similar estimators) sharply reduces the impact of a small fraction of arbitrarily corrupted ratings, but once the dominant failure mode is a shared latent shift (high ICC), all estimators inherit the $\sqrt{\rho}$ plateau. This matches the feasibility condition $\rho \lesssim \tau^2$ above.

(2) "More raters" exhibits diminishing returns when $\rho > 0$. Even in the absence of explicit contamination, the marginal gain from increasing m drops quickly as m exceeds $O(1/\rho)$. This is the numerical face of $\chi(m) = 1 + (m-1)\rho$: the effective sample size saturates near $1/\rho$.

(3) Diversification acts like a structural reduction in χ. If we simulate stratified panels where ratings are drawn from multiple pools with low cross-pool correlation (e.g. each block of a median-of-means estimator draws one rater per pool), the empirical concentration tracks a *block-level* design effect rather than the within-pool ICC. This suggests that how we allocate raters to blocks can be as important as the total m.

Operationally, these simulations let us produce domain-specific "audit sizing curves": for each query class, a map from (τ, δ) to the minimal m achieving an empirical failure probability below δ under conservative (ε, ρ) assumptions. These curves are the actionable bridge from the theoretical audit bound to a budget and staffing plan.

8.5 What we should report before claiming "constant-cost"

Before moving to oversight design, we can state a minimal empirical bar for making the constant-cost claim credible in a domain: we should be able to (i) bound an effective ICC ρ (or directly χ) for the intended panel design, (ii) bound a contamination-like bias floor ε via red-teaming and held-out gold queries, and (iii) demonstrate by simulation (or live fire drills) that the chosen robust estimator with the chosen m achieves the intended local tolerance $\tau = \Theta(1/K)$ at the target δ. The next section uses these requirements to derive concrete procurement and process implications: how to engineer independence, how to standardize robust aggregation, and how to integrate calibrated audits into training and governance workflows.

9 Implications for oversight design: procurement, aggregation standards, and training integration

The calibration story in Section 8 has a concrete operational moral: if we want "single-audit" oversight to scale, then we must treat *independence* and *bias control* as first-class engineered resources. In our formalism, these resources appear as the effective dependence inflation χ and the contamination rate ε. In deployment, however, χ and ε are not abstract constants; they are functions of procurement (who we hire), workflow (how we isolate and incentivize them), and interface (what context we show). This section translates the theorem's knobs into process requirements, and flags the main failure modes when those requirements are not met.

9.1 Panel procurement as an independence engineering problem

The most important design lesson is that "more raters" is not the same as "more information." Once correlation is present, the marginal value of additional ratings can collapse, and the relevant objective becomes minimizing χ subject to budget and latency constraints. Practically, this pushes us toward *diversified panels* rather than larger homogeneous pools.

A procurement pipeline that targets low χ typically includes:

- **Multiple pools with low shared context.** We want raters whose errors are not driven by a single shared artifact (one training document, one subreddit, one internal style guide). Using multiple vendors, geographies, and professional backgrounds is not "fairness theater" here; it is a concentration bound.

- **Isolation and non-communication controls.** If raters can coordinate (or are exposed to the same justifications generated by the model), correlation increases and may become *strategic*. Operationally, this means time-windowed tasks, no shared chat, and strict controls on what model-produced artifacts are shown.

- **Prompt randomization and counterbalancing.** Small framing changes can dominate outcomes for ambiguous queries; randomizing presentation order, paraphrasing, and hiding irrelevant metadata is often the cheapest way to reduce common shocks.

- **Rater rotation and freshness.** Repeated exposure can create shared heuristics and brittle shortcuts. Rotating raters across tasks and keeping a "fresh" reserve pool reduces the chance that χ drifts upward over time.

These steps also interact with ε. In our model, ε captures arbitrary contamination: adversarial raters, compromised accounts, or persistent systematic biases not covered by the benign dependence model. Procurement must therefore include adversarial thinking: background checks where appropriate, rate-limiting and anomaly detection, and *continuous* red-teaming to estimate a conservative $\widehat{\varepsilon}$ (e.g., via hidden gold questions and canary tasks). A key governance takeaway is that if we cannot bound ε away from the information-theoretic limit (e.g. $\varepsilon \approx 1/2$), then no amount of clever aggregation will save a constant-cost audit.

Finally, we can exploit the structure of robust estimators when allocating raters. Median-of-means (MoM) partitions ratings into blocks; if we can assign raters so that each block draws from distinct pools, then the relevant dependence becomes closer to a *block-level* design effect. Put differently: the procurement goal is not merely "low within-pool ICC," but "low cross-block correlation," because the protocol's test statistic is block-robust.

9.2 Robust aggregation standards as a safety interface

Once we accept that audits are statistical hypothesis tests under contamination and dependence, we should standardize the estimator and its reporting in the same way safety-critical fields standardize test procedures. The verifier's decision rule depends on the event

$$|\tilde{p}_t - \hat{p}_t| \geq \tau,$$

so the aggregation pipeline should specify, at minimum: (i) how $\tilde{p}_t$ is computed, (ii) how blocks are formed, (iii) how missing or low-quality ratings are handled, and (iv) what auxiliary diagnostics are logged.

A reasonable "default standard" consistent with our guarantees is:

- **Estimator:** MoM with $g \asymp \log(2/\delta)$ blocks, fixed publicly by policy, and block size $b = m/g$ chosen to satisfy the constant requirements of the robust mean lemma.
- **Block construction:** stratified by pool where possible (one rater per pool per block), and randomized using an auditable randomness source to prevent either prover from steering which raters land in which block.
- **Robustness checks:** log the empirical block means and their dispersion; a high between-block spread is an early warning for either contamination (large ε) or shared ambiguity (high χ). These diagnostics are also the raw material for updating $\widehat{\chi}$ online.

We emphasize that *standardization is part of the security model.* If different teams choose different aggregators ad hoc, the system creates attack surface: the agent can learn which variants are easiest to manipulate and route tasks accordingly. In economic terms, aggregation is a commitment device that shrinks the strategic space.

There is also a subtle but important deployment constraint: robust estimation mitigates *outliers*, not *shared shifts*. If a query template systematically induces a common wrong intuition (high ICC), then MoM will confidently return the wrong answer. This is exactly the regime where the feasibility condition $\rho \lesssim \tau^2$ fails. Operationally, the remedy is not "increase m," but "change the prompt / panel design" or "escalate" (e.g. expert review, multi-step deliberation, or abandoning single-audit for that task class).

9.3 Integrating calibrated audits into reward models and training loops

A natural next step is to use audited outcomes not only for one-off verification, but also as training signal—for example, training a reward model (RM) to predict $\tilde{p}_t$, or training the agent to propose actions that pass audits. This

can reduce audit frequency over time, but it also introduces new feedback loops.

From our perspective, the key principle is: *audits are a stochastic oracle with known failure modes.* If we train directly on audit outcomes without preserving independence, we risk "teaching to the test": the model adapts to rater quirks, increasing effective χ (shared exploitation) and potentially increasing ε (if compromised raters become focal). Concretely, we recommend:

- **Hold-out oversight:** maintain a separate, rotating hold-out panel design for evaluation and for the occasional "ground truth" recalibration of $(\widehat{\varepsilon}, \widehat{\chi})$.
- **Distribution shift monitoring:** as the model improves, it may seek edge cases where humans correlate strongly. Monitoring the time series of ICC proxies and block-dispersion diagnostics can detect this.
- **Conservative label use:** treat $\tilde{p}_t$ as a noisy label with a bias floor. In particular, if the theoretical error is $\tilde{O}(\sqrt{\chi/m} + \varepsilon)$, then downstream learners should incorporate this as label uncertainty, rather than overfitting.

A useful framing is to view audits as scarce, high-integrity data used to *calibrate* cheaper automated checks. In governance terms, this suggests a two-layer control system: automated monitoring for breadth, and calibrated human audits for depth, with clear triggers for escalation when the automated layer drifts.

9.4 Open problems and likely breaking points

Several gaps remain between the clean theorem statement and a robust end-to-end oversight system.

Strategic dependence and endogenous χ. Our dependence control is exogenous; in reality, a capable model may actively *create* correlation (e.g. by generating misleading but compelling rationales that many raters adopt). Understanding protocols where χ is adversarially influenced is a priority.

Online estimation with incentives. We sketched pilot-based estimation of ICC and contamination, but deployments are nonstationary. We need sequential methods that update $\widehat{\chi}$ and $\widehat{\varepsilon}$ under strategic behavior, while preventing Goodharting against the estimator itself.

Beyond single-audit: adaptive and multi-step audits. Single-audit protocols are attractive because they bound human cost, but they are brittle when ρ is large or when K forces τ to be small. Characterizing the optimal

tradeoff between "one expensive audit" and "a few cheap audits" (and how this composes with the debate transcript) remains open.

Task-dependent Lipschitzness and interface design. The parameter K is a property of the machine and oracle interface, not just the task. Designing oversight query languages that reduce effective sensitivity (smaller K) could be as impactful as improving panels.

Overall, the safety tradeoff is now explicit: constant-cost oversight is feasible only if we can *buy* sufficiently low χ and ε through procurement and process. Where we cannot, the correct response is not to overclaim theoretical scalability, but to switch regimes: increase audit depth, change the oracle interface, or restrict deployment to domains where calibrated independence is achievable.