

Can a Small Audit Certify Rare Online Harms? Finite-Sample Limits and Prediction-Assisted Prevalence Bounds

Anonymous submission

Abstract

External audits are increasingly deployed to measure the prevalence of harmful content on online platforms and certify compliance with audit thresholds. When target harms are rare ($\pi \approx 10^{-3}$), however, conventional audits face severe, under-appreciated statistical limits. Under realistic human-label budgets ($n \in [500, 2,000]$), we show that standard audits suffer a strict finite-sample *resolution ceiling*: observing zero harmful items out of $n = 1,000$ random samples only bounds prevalence below 0.30% at 95% confidence ($U_{95\%} \approx 3/n$), rendering claims of compliance with rare-harm thresholds statistically unsupported. While practitioners frequently turn to model extrapolation by transferring historical classifier calibration to unlabelled target periods, we evaluate historical stable-calibration extrapolation (fitting non-parametric isotonic calibration on historical data and extrapolating over the target census without fresh target annotations) and demonstrate that uncorrected temporal and community calibration drift collapses empirical coverage from 95.4% (in stationary data) to 55.8% across 500 complete calibration resamples ($n_{\text{cal}} = 5,000$, $n = 500$), refitting the isotonic model on each calibration draw. To resolve this dilemma, we instantiate finite-population prediction-powered inference (PPI) using a one-sided without-replacement betting bound on bounded prediction residuals. By coupling uncalibrated auxiliary classifier scores across the census with a small fresh probability sample, this approach guarantees non-asymptotic $\geq 95\%$ coverage for any fixed target population and auxiliary score vector under simple random sampling without replacement, without assumptions on historical-to-target distribution shift, while delivering a $2.6\times$ to $3.7\times$ reduction in required human labels for equivalent 80% certification power across the four primary harms in the Civil Comments ecosystem (Threat at $\pi = 0.12\%$, Severe Toxicity at $\pi = 0.15\%$, Sexually Explicit content at $\pi = 0.28\%$, and Identity Attack at $\pi = 0.42\%$; e.g., $1,255 \rightarrow 485$ for Threat at $\delta = 0.50\%$, and $1,575 \rightarrow 455$ for Sexually Explicit at $\delta = 0.75\%$). Finally, we reveal an *aggregation masking* effect: aggregate platform-wide threat prevalence falls below an illustrative audit threshold of 0.20%, yet this conceals the fact that 19.4% of localized publication community-quarter cells exceed that threshold (reaching up to 0.48%). We translate these findings into concrete operational recommendations for platform auditors and regulatory bodies.

1 Introduction

Audits of web-scale communication platforms and generative foundation models increasingly serve as the quantitative benchmark for safety compliance (Raji et al. 2020; Raghavan et al. 2020). Regulatory frameworks such as the European Union Digital Services Act (DSA) and emerging international governance standards mandate third-party measurement of systemic risks, including violent threats, hate speech, severe toxicity, and sexually explicit text. In platform transparency reports and regulatory filings, online services frequently declare that the prevalence of harmful content has been curtailed below specified safety benchmarks, such as 0.10% or 0.05% of total content impressions.

However, auditing extremely rare online harms introduces severe statistical obstacles that standard evaluation paradigms fail to address. External auditors typically operate under stringent human-annotation budgets, inspecting between several hundred and a few thousand items ($n \in [500, 2,000]$) (Torabi et al. 2020; Wu and Resnick 2024). When the true prevalence π of a severe violation is exceedingly low (e.g., $\pi = 0.12\%$), a standard random sample of $n = 1,000$ items will, with high probability ($(1 - \pi)^n \approx 30.1\%$), contain *zero* positive instances.

A common but dangerous intuition treats zero observed violations as decisive evidence that the platform is compliant. Inverting the exact hypergeometric or binomial distribution reveals the fallacy of this interpretation: observing zero violations in 1,000 samples yields a 95% one-sided upper confidence bound of $U_{95\%} \approx 0.30\%$. Consequently, an audit of 1,000 items is mathematically incapable of certifying that prevalence is below 0.10%, no matter how clean the sample appears. Standard large-sample normal approximations ($z\sqrt{\hat{p}(1 - \hat{p})/n}$) fail completely in this sparse regime, yielding a degenerate standard error of zero and an impossible upper bound of zero.

To overcome this resolution limit without inflating human-annotation costs, recent literature has explored model-assisted estimation. A prominent line of work relies on *historical stable-calibration extrapolation* procedures (Wu and Resnick 2024; Kumar, Liang, and Ma 2019), which learn a post-hoc calibration mapping $\hat{g}(q) \approx P_{\text{hist}}(Y = 1 \mid q)$ relating automated classifier scores q to human judgments on historical data, and subsequently extrapolate this function to estimate target-period prevalence without fresh target anno-

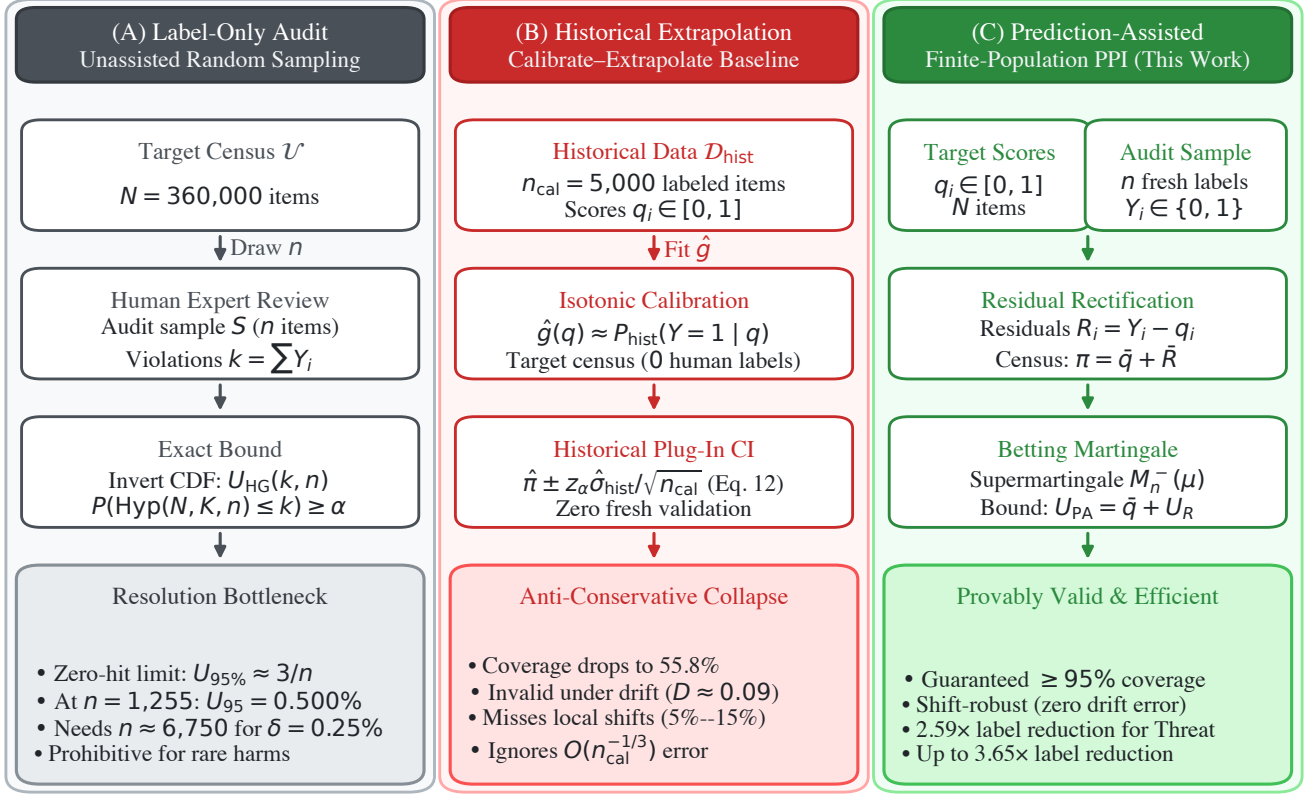


Figure 1: **Conceptual comparison of rare-harm auditing paradigms.** (A) **Label-Only Audit:** Inverting exact hypergeometric bounds yields a resolution limit of $U_{95\%} \approx 3/n$ when zero hits are observed, requiring prohibitive human-label budgets ($n > 3,000$) to certify prevalence below 0.10%. (B) **Historical Stable-Calibration Extrapolation:** Post-hoc calibration functions fit on historical data produce narrow intervals, but become severely anti-conservative under temporal or community calibration drift, causing empirical coverage to collapse from 95.4% to 55.8% across 500 complete calibration resamples ($n_{\text{cal}} = 5,000$, refitting calibration per draw, $n = 500$). (C) **Prediction-Assisted Rectification (Finite-Population PPI):** By coupling uncalibrated auxiliary model predictions with a fresh probability sample via without-replacement betting martingales, our approach guarantees non-asymptotic $\geq 95\%$ coverage for any fixed target census without distribution shift assumptions, while achieving 2.6× to 3.7× human-label efficiency gains.

tations. While historical calibration extrapolation produces narrow confidence intervals, its validity rests entirely on a strong invariance assumption: that the conditional probability of human harm given classifier scores remains stationary over time and across distinct sub-communities. As Wu and Resnick (2024) emphasize, establishing the validity of this stability assumption is essential. In fast-evolving social environments, language use, adversarial evasion, and moderation dynamics continuously alter the relationship between model predictions and human harm. When this assumption is violated, extrapolation becomes anti-conservative, producing falsely tight bounds that fail to cover the true target prevalence.

In this work, we investigate four fundamental research questions:

- **RQ1 (Audit Resolution Limits):** At realistic human-label budgets, how small a harmful-content prevalence can a conventional probability audit rule out with 95% confidence, and what are the exact sample complexity requirements?

- **RQ2 (Calibration Drift Fragility):** How sensitive is historical stable-calibration extrapolation to temporal and cross-community distribution shifts, and under what conditions do uncertainty bounds collapse?
- **RQ3 (Prediction-Assisted Rectification):** Can finite-population prediction-powered inference (PPI) provide valid, non-asymptotic one-sided certificates for rare harms without assuming classifier calibration or temporal stability, and what sample efficiency gains does it deliver?
- **RQ4 (Aggregation Masking):** When an audit evaluates compliance at the platform aggregate level, how much localized harm can be masked across individual communities and time periods?

Contributions. We clarify that we do not claim the statistical difference estimator or without-replacement betting martingales as novel inventions; the foundational constructions originate in model-assisted survey sampling (Särndal, Swensson, and Wretman 2003; Cochran 1977), prediction-powered inference (Angelopoulos et al. 2023), design-

based supervised learning (Egami et al. 2023), and without-replacement betting martingales (Waudby-Smith and Ramdas 2024). Instead, our contributions lie in formalizing, analyzing, and resolving the statistical dilemma of auditing rare online harms:

1. **Finite-Sample Resolution Characterization (RQ1):** We prove the finite-sample resolution ceiling for rare-harm audits under exact binomial and hypergeometric sampling without replacement. We demonstrate that observing zero violations out of $n = 1,000$ items yields $U_{95\%} \approx 3/n = 0.30\%$, and derive exact sample sizes required for rare-event certification.
2. **Quantifying Calibration Drift Collapse (RQ2):** Through extensive simulations on the Civil Comments ecosystem, we quantify the empirical risk of historical stable-calibration extrapolation under real-world temporal and community shifts. Across 500 complete calibration resamples ($n_{\text{cal}} = 5,000$, refitting isotonic calibration on each draw, zero fresh target labels), empirical coverage plummets from 95.4% in stationary data to 55.8% under calibration drift D_{cal} ($n = 500$), producing anti-conservative certificates.
3. **Instantiating Finite-Population PPI for Rare Harms (RQ3):** We instantiate finite-population prediction-assisted inference for one-sided rare-harm certification using a one-sided without-replacement betting bound (specializing Waudby-Smith & Ramdas, 2024) on bounded residuals. We prove that this approach guarantees non-asymptotic $\geq 95\%$ coverage for any fixed target population and score vector without requiring classifier calibration or drift assumptions, while reducing human-label budgets by $2.6\times$ to $3.7\times$ for 80% certification power across the four primary harms.
4. **Demonstrating Aggregation Masking (RQ4):** We uncover a severe aggregation masking effect across 42 publication communities over six calendar quarters: an aggregate platform threat prevalence of 0.12% falls well below an illustrative audit threshold of 0.20%, yet conceals that 19.4% of localized community-quarter cells exceed the threshold (reaching up to 0.48%). We synthesize these findings into five concrete operational guidelines for independent auditors and regulators.

2 Related Work

Algorithmic Auditing and Content Moderation. External and internal audits are foundational to algorithmic accountability and platform governance (Raji et al. 2020; Raghavan et al. 2020). Regulatory regimes, notably the EU Digital Services Act, require very large online platforms (VLOPs) to facilitate third-party risk assessments of systemic societal harms, such as hate speech, illegal incitement, and child sexual abuse material. Standard industry auditing practice relies on manual review of small random samples or stratified cohorts (Torabi et al. 2020). However, as highlighted in recent measurement studies, auditing rare events with small samples suffers from extreme variance and vulnerability to selection bias (Wu and Resnick 2024).

Prevalence Estimation and Calibration Extrapolation.

When human labeling is expensive, automated classifiers are frequently used as measurement surrogates. To correct for classification errors, early methods adjusted raw model classifications using confusion matrices or inversion formulas (Zadrozny and Elkan 2002). In modern neural NLP pipelines, post-hoc calibration techniques—including Platt scaling (Platt 1999), isotonic regression (Zadrozny and Elkan 2002), and temperature scaling (Guo et al. 2017)—map uncalibrated logits to pseudo-probabilities. Wu and Resnick (2024) systematically investigated whether historical classifier calibration can be extrapolated to measure prevalence over subsequent time periods in online communication. Crucially, Wu and Resnick (2024) recognized that extrapolation relies heavily on the stability of the conditional distribution $P(Y | q)$ across contexts. When distribution shift or concept drift occurs (Tibshirani et al. 2019; Gupta, Podkopaev, and Ramdas 2020), calibrated estimates become biased. Torabi et al. (2020) proposed Bucketed Beta-Binomial (BBB) models to monitor rare violations, leveraging historical score buckets as informative priors. However, BBB models remain sensitive to prior misspecification under localized distribution shifts.

Prediction-Powered Inference and Design-Based Learning.

Our framework builds on the rich intersection of survey sampling and statistical inference with machine learning predictions. Classical model-assisted survey sampling (Särndal, Swensson, and Wretman 2003; Cochran 1977) developed difference estimators to improve the precision of finite-population means using auxiliary regression models. Recently, Angelopoulos et al. (2023) generalized this concept into Prediction-Powered Inference (PPI), enabling valid statistical inference using machine learning predictions. Appendices B and E of Angelopoulos et al. (2023) established finite-population mean estimation via residual confidence intervals using without-replacement concentration bounds (Waudby-Smith and Ramdas 2024). In political science, Egami et al. (2023) independently established Design-based Supervised Learning (DSL) for downstream causal and descriptive inference.

While existing PPI literature predominantly focuses on two-sided asymptotic normal confidence intervals for moderate-prevalence settings, rare online harms ($\pi \leq 0.50\%$) violate standard asymptotic normality assumptions under small sample sizes. By leveraging finite-population without-replacement betting martingales (Waudby-Smith and Ramdas 2024; Bardenet and Maillard 2015; Ville 1939), we instantiate a non-asymptotic one-sided certification framework tailored specifically to sparse harmful-content audits.

3 Problem Formulation & Audit Limits (RQ1)

3.1 Finite-Population Setup and Certification Objective

Let $U = \{1, 2, \dots, N\}$ denote a finite census of online content items (e.g., all comments, posts, or messages published

during a given audit window). For each item $i \in U$, let $Y_i \in \{0, 1\}$ represent the true, unobserved human harm label, where $Y_i = 1$ indicates a policy violation (e.g., violent threat, severe toxicity) and $Y_i = 0$ indicates non-violating content. The target quantity of interest is the true population prevalence:

$$\pi = \frac{1}{N} \sum_{i=1}^N Y_i. \quad (1)$$

In regulatory and compliance audits, an auditor does not merely seek an unconstrained point estimate of π ; rather, the operational objective is to certify whether the platform complies with a prespecified safety threshold $\delta \in (0, 1)$. Formally, the auditor tests the null hypothesis of non-compliance against the alternative of certified compliance:

$$H_0 : \pi \geq \delta \quad \text{versus} \quad H_1 : \pi < \delta. \quad (2)$$

A valid audit procedure with significance level $\alpha \in (0, 1)$ constructs a data-dependent upper confidence bound $U_{1-\alpha}$ such that:

$$P(\pi \leq U_{1-\alpha}) \geq 1 - \alpha. \quad (3)$$

If $U_{1-\alpha} \leq \delta$, the auditor rejects H_0 and certifies that harmful-content prevalence does not exceed δ with at least $(1 - \alpha)$ confidence.

3.2 The Zero-Hit Fallacy and Finite-Sample Resolution Limits

In a conventional audit, the auditor draws a simple random sample without replacement $S \subset U$ of size $n = |S| \ll N$. Human expert annotators label every item in S , observing $k = \sum_{i \in S} Y_i$ violations. The sample proportion is $\hat{\pi} = k/n$.

When auditing rare harms ($\pi \leq 0.20\%$) under standard audit budgets ($n \in [500, 2,000]$), the probability of observing zero violations in the sample is substantial:

$$P(k = 0) = \frac{\binom{N-N\pi}{n}}{\binom{N}{n}} \approx (1 - \pi)^n \approx e^{-n\pi}. \quad (4)$$

For example, if $\pi = 0.12\%$ and $n = 1,000$, $P(k = 0) \approx e^{-1.2} \approx 30.1\%$.

A naive practitioner might attempt to apply standard asymptotic normal confidence intervals:

$$U_{\text{asympt}} = \hat{\pi} + z_{1-\alpha} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n} \left(\frac{N - n}{N - 1} \right)}. \quad (5)$$

When $k = 0$, $\hat{\pi} = 0$, yielding an estimated standard error of zero and $U_{\text{asympt}} = 0$. This gives an absurd and statistically invalid claim that harmful content is completely absent from the platform ($\pi = 0$).

To maintain mathematical validity, the auditor must invert the exact finite-population hypergeometric distribution or its binomial envelope via the Clopper–Pearson method (Clopper and Pearson 1934).

Proposition 1 (Exact Hypergeometric Inversion and Zero-Hit Resolution Limit). *Let S be a simple random sample of size n drawn without replacement from a finite population of size N , containing an unknown number of violations $K = N\pi$. Let $k = \sum_{i \in S} Y_i$ denote the observed violations.*

The exact $1 - \alpha$ one-sided finite-population hypergeometric upper confidence bound U_{HG} is:

$$U_{\text{HG}} = \frac{1}{N} \max_{K \leq N} \{K : P(X \leq k \mid N, K, n) \geq \alpha\}, \quad (6)$$

where $X \sim \text{Hypergeometric}(N, K, n)$. When zero violations are observed ($k = 0$) and the sampling fraction is small ($n/N \ll 1$), the large- N limit satisfies:

$$U_{1-\alpha} = 1 - \alpha^{1/n} \approx \frac{\ln(1/\alpha)}{n}. \quad (7)$$

For nominal confidence 95% ($\alpha = 0.05$), since $\ln(20) \approx 2.9957 \approx 3$, the finite-sample resolution limit is:

$$U_{95\%} \approx \frac{3}{n}. \quad (8)$$

Proof. Under simple random sampling without replacement of size n from population N with K positive items, the number of observed violations X follows the hypergeometric distribution $P(X \leq k \mid N, K, n) = \sum_{j=0}^k \binom{K}{j} \binom{N-K}{n-j} / \binom{N}{n}$. The exact $1 - \alpha$ upper confidence bound K_{max} inverts the one-sided hypothesis test $H_0 : K \geq K_0$, accepting all K for which $P(X \leq k) \geq \alpha$, yielding $U_{\text{HG}} = K_{\text{max}}/N$. In the large- N limit ($n/N \ll 1$), sampling without replacement converges to sampling with replacement with success parameter $p = K/N$. When zero successes are observed ($k = 0$), the probability of zero hits is $P(k = 0) = (1 - p)^n$. Inverting the test $P_p(k \leq 0) = \alpha$ yields $(1 - U_{1-\alpha})^n = \alpha$, whence $U_{1-\alpha} = 1 - \alpha^{1/n}$. Using the first-order Taylor expansion $-\ln(1 - U) \approx U$ for $U \ll 1$, we obtain $U_{1-\alpha} \approx \frac{-\ln \alpha}{n} = \frac{\ln(1/\alpha)}{n}$. Setting $\alpha = 0.05$ yields $\ln(20) \approx 2.996 \approx 3$, completing the proof. $\square$

Implications for Rare-Harm Certification and Power Mechanics. Proposition 1 establishes a fundamental resolution floor for label-only audits:

- At $n = 500$: $U_{95\%} \approx 3/500 = 0.60\%$.
- At $n = 1,000$: $U_{95\%} \approx 3/1,000 = 0.30\%$.
- At $n = 2,000$: $U_{95\%} \approx 3/2,000 = 0.15\%$.

To certify that prevalence is below $\delta = 0.10\%$ under a clean audit ($k = 0$), an auditor must review at least $n_{\text{req}} \geq \frac{\ln(20)}{0.0010} \approx 2,996$ items. If violations are observed, inverting the exact hypergeometric distribution yields $U_{95\%} \approx 4.74/n$ for $k = 1$ ($n \geq 4,744$ for $\delta = 0.10\%$) and $U_{95\%} \approx 6.30/n$ for $k = 2$ ($n \geq 6,296$).

Crucially, an audit certifies compliance at threshold δ if and only if $U_{\text{HG}}(k, n) \leq \delta$, defining a critical maximum allowable hit count $k^*(n, \delta) = \max\{k \geq 0 : U_{\text{HG}}(k, n) \leq \delta\}$. The certification power under true prevalence π is the probability of observing at most $k^*(n, \delta)$ hits:

$$\text{Power}(n, \delta, \pi) = P(X \leq k^*(n, \delta)), \quad (9)$$

where $X \sim \text{Hypergeometric}(N, N\pi, n)$.

For Threat ($\pi = 0.12\%$, $N = 360,000$, true count $K = 432$):

- When threshold is set to $\delta = 0.50\%$ (slack $+0.38\%$), exact hypergeometric inversion yields $U_{\text{HG}}(2, 1255) = 0.5000\% \leq \delta$, while at $n = 1,254$, $U_{\text{HG}}(2, 1254) =$

0.5004% > δ (where only $k \leq 1$ certifies, giving power 55.6%). Thus, at $n = 1,255$, allowable hits include $k \leq 2$, achieving exact hypergeometric power $P(X \leq 2) = 80.76\% \approx 80.8\%$.

- Conversely, if an auditor set an aggressive threshold of $\delta = 0.25\%$ (slack +0.13%), exact hypergeometric inversion requires $n = 6,748 \approx 6,750$ human labels to achieve 80% power ($k^* = 10$, $U_{\text{HG}} = 0.2500\%$, power = 80.80%). At $n = 1,255$, only $k = 0$ certifies ($U_{\text{HG}}(0) = 0.239\% \leq 0.25\%$, whereas $k = 1$ yields $0.378\% > 0.25\%$), which under $\pi = 0.12\%$ yields certification power of only $P(k = 0) \approx 22.1\%$. At $n = 7,400$, the critical allowable count is $k^* = 11$ ($U_{\text{HG}} = 0.2447\% \leq 0.25\%$), yielding $81.67\% \approx 81.7\%$ power.

Thus, standard label-only audits are constrained by a strict trade-off between threshold stringency and human annotation budgets.

4 Historical Stable-Calibration Extrapolation Under Drift (RQ2)

To bypass the severe sample size requirements of human annotation, platforms and researchers frequently turn to machine learning models as measurement surrogates (Wu and Resnick 2024; Torabi et al. 2020). Suppose an automated content moderation classifier produces a continuous violation score $q_i \in [0, 1]$ for every item $i \in U$. Because raw model predictions are typically miscalibrated, practitioners employ the *historical stable-calibration extrapolation* workflow, predominantly using non-parametric isotonic regression (Wu and Resnick 2024).

4.1 The Historical Calibration Extrapolation Baseline

Historical calibration extrapolation proceeds in three steps: (1) **Historical Calibration:** On a historical labeled dataset $D_{\text{hist}} = \{(q_j, Y_j)\}_{j=1}^{n_{\text{cal}}}$, fit a non-parametric isotonic calibration mapping $\hat{g} : [0, 1] \rightarrow [0, 1]$ via the Pool Adjacent Violators (PAV) algorithm:

$$\hat{g}(q) \approx P_{\text{hist}}(Y = 1 \mid q). \quad (10)$$

(2) **Census Extrapolation:** Compute model scores across the unlabelled target census U_{target} and evaluate the calibrated scores:

$$\hat{\pi}_{\text{CE}} = \frac{1}{N} \sum_{i \in U_{\text{target}}} \hat{g}(q_i). \quad (11)$$

(3) **Uncertainty Quantification:** While Wu and Resnick (2024) originally quantified uncertainty via bootstrap resampling of calibration draws, an analytical simplification often evaluated in practice is the *plug-in historical-extrapolation bound*, which treats $\hat{g}$ as fixed:

$$U_{\text{CE}}^{\text{plug}} = \hat{\pi}_{\text{CE}} + z_{1-\alpha} \sqrt{\frac{\widehat{\text{Var}}_{\text{hist}}(\hat{g}(q))}{n_{\text{cal}}}}. \quad (12)$$

As noted by Wu and Resnick (2024), Eq. 12 ignores the estimation variability of the non-parametric isotonic estimator $\hat{g}$, which converges at the non-standard $n_{\text{cal}}^{-1/3}$ rate.

However, even when uncertainty is propagated using non-parametric calibration bootstrap ($B = 200$ resamples) as proposed by Wu and Resnick (2024), coverage still plummets under distribution shift. The bootstrap resamples exclusively from historical calibration pairs $(q_j, Y_j) \sim P_{\text{hist}}$, capturing sampling variability around the historical expectation $\mathbb{E}_{P_{\text{hist}}}[\hat{g}(q)]$. It cannot detect or expand for the target domain shift $\mathbb{E}_{P_{\text{target}}}[\hat{g}(q)] \neq \pi_{\text{target}}$, leaving the underlying drift bias completely uncorrected.

4.2 Distribution Shift and Coverage Collapse

As Wu and Resnick (2024) observe, the validity of $\hat{\pi}_{\text{CE}}$ hinges on a critical transportability assumption:

$$P_{\text{target}}(Y = 1 \mid q) = P_{\text{hist}}(Y = 1 \mid q) \quad \forall q \in [0, 1]. \quad (13)$$

In production social platforms, this assumption is routinely violated by temporal drift (evolving language, events, evasion) and community heterogeneity (sub-communities with distinct base rates and norms).

To quantify calibration divergence from finite data, we estimate historical and target calibration functions $\hat{g}_{\text{hist}}(q)$ and $\hat{g}_{\text{target}}(q)$ via isotonic regression on historical calibration comments ($n_{\text{cal}} = 5,000$) and target comments ($N = 360,000$). Isotonic regression naturally accommodates low-density score regions ($q > 0.50$) by pooling adjacent sparse intervals into monotonic steps without requiring arbitrary kernel bandwidth tuning. We define the *Calibration Drift Metric* D_{cal} as the unweighted integrated absolute deviation across score space:

$$D_{\text{cal}} = \int_0^1 |\hat{g}_{\text{target}}(q) - \hat{g}_{\text{hist}}(q)| dq, \quad (14)$$

evaluated numerically over a fine uniform grid of 1,000 points. Across 500 calibration resamples, D_{cal} averages 0.091 ± 0.008 (95% range $[0.076, 0.108]$), indicating a 9.1 percentage-point mean vertical shift across calibration curves.

Units and Distinction from Aggregate Prevalence Bias.

We emphasize the distinction between D_{cal} and aggregate prevalence bias Δ_{bias} . $D_{\text{cal}} \approx 0.091$ measures the unweighted curve distance across score space $[0, 1]$. Because rare-harm violations concentrate in the upper score tail, local calibration shifts of 0.05 to 0.15 in $q \geq 0.10$ drive this curve divergence. When weighted by the empirical target density $P_{\text{target}}(q)$ (where $> 98\%$ of non-violating items have $q \approx 0$), this curve shift translates into a population-weighted prevalence bias:

$$\Delta_{\text{bias}} = \frac{1}{N} \sum_{i=1}^N (\hat{g}_{\text{hist}}(q_i) - P_{\text{target}}(Y_i = 1 \mid q_i)). \quad (15)$$

Empirically, $\Delta_{\text{bias}} \approx 0.06\%$. Crucially, while $\Delta_{\text{bias}} \approx 0.06\%$ is small in absolute probability units, it constitutes a massive 50% relative distortion against a baseline prevalence of $\pi = 0.12\%$. This bias overwhelms confidence widths, collapsing empirical coverage of the plug-in bound to 55.8% and the calibration bootstrap to 18.5% (underestimating prevalence in 81.5% of audits).

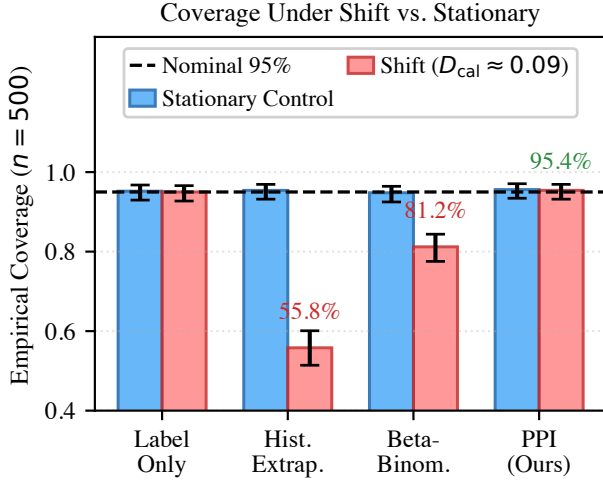


Figure 2: **Empirical coverage under stationary vs. shifted calibration regimes** ($n = 500$, Threat harm). Under the stationary control ($D_{\text{cal}} = 0$), historical extrapolation attains approximately nominal 95.4% coverage across 500 calibration resamples ($n_{\text{cal}} = 5,000$). However, under realistic temporal and cross-community calibration drift ($D_{\text{cal}} \approx 0.091$), historical stable-calibration extrapolation collapses catastrophically to 55.8% (and bootstrap-aware extrapolation to 18.5%), and Bucketed Beta-Binomial degrades to 81.2%. In contrast, Prediction-Assisted rectification (96.8% $\rightarrow$ 95.8%) and Label-Only audits (99.2% $\rightarrow$ 99.0%) strictly maintain valid $\geq 95\%$ coverage across both regimes. Error bars denote 95% Wilson score confidence intervals.

Bucketed Beta-Binomial (BBB) Limitations. Torabi et al. (2020) proposed partitioning model scores into B discrete buckets and placing Beta priors on bucket-level hit rates: $\theta_b \sim \text{Beta}(\alpha_b, \beta_b)$. Model scores are partitioned into $B = 5$ quantile-based strata with boundaries $\{0.00, 0.01, 0.05, 0.20, 0.50, 1.00\}$ and census weights $W_b = N_b/N$. For each stratum b , historical calibration data yields $(y_b^{\text{cal}}, n_b^{\text{cal}})$, defining conjugate prior $\text{Beta}(\alpha_b, \beta_b)$ with effective sample size $n_{\text{prior}} = 40.5$ ($\alpha_b = 0.5 + \kappa y_b^{\text{cal}}, \beta_b = 40.0 + \kappa(n_b^{\text{cal}} - y_b^{\text{cal}})$ where $\kappa = 40.5/n_b^{\text{cal}}$). In the audit sample of size n , observed counts (y_b, n_b) yield posterior $\text{Beta}(\alpha_b + y_b, \beta_b + n_b - y_b)$; the 95% upper bound U_{BBB} is computed as the 95th percentile of the stratified mixture $\sum_{b=1}^B W_b \theta_b$. Under drift, the historical prior regularizes the posterior toward stale hit rates, causing undercoverage (81.2% at $n = 500$, degrading to 71.8% at $n = 2,000$ in Table 1).

5 Prediction-Assisted Prevalence Bounds (RQ3)

To overcome the finite-sample resolution limit of label-only audits without falling victim to calibration drift, we instantiate finite-population prediction-powered inference (PPI) (Angelopoulos et al. 2023) using without-replacement betting martingales (Waudby-Smith and Ramdas 2024).

5.1 Finite-Population Prediction-Assisted Difference Identity

Suppose the auditor has access to uncalibrated auxiliary model predictions $q_i \in [0, 1]$ for every item i in the census U . Crucially, we make *no* assumptions about how q_i was generated: the model may be poorly calibrated, biased, or trained on historical data from completely different distributions.

The true population prevalence can be decomposed identically as:

$$\pi = \frac{1}{N} \sum_{i=1}^N Y_i = \underbrace{\frac{1}{N} \sum_{i=1}^N q_i}_{\bar{q}} + \underbrace{\frac{1}{N} \sum_{i=1}^N (Y_i - q_i)}_{\bar{R}}. \quad (16)$$

Here:

- $\bar{q} = \frac{1}{N} \sum_{i=1}^N q_i$ is the *census prediction mean*, which is deterministic and known exactly with zero sampling error by running the classifier over U .
- $R_i = Y_i - q_i \in [-1, 1]$ is the prediction residual for item i , and $\bar{R} = \frac{1}{N} \sum_{i=1}^N R_i$ is the true finite-population residual mean.

Because $\bar{q}$ is known, bounding the unknown prevalence π is exactly equivalent to bounding the finite-population mean $\bar{R}$ of bounded variables $R_i \in [-1, 1]$. To map directly to the canonical without-replacement betting martingale framework of Waudby-Smith and Ramdas (2024), we apply the linear transformation:

$$Z_i = \frac{R_i + 1}{2} = \frac{Y_i - q_i + 1}{2} \in [0, 1]. \quad (17)$$

The population mean of Z satisfies:

$$\mu_Z = \frac{1}{N} \sum_{i=1}^N Z_i = \frac{\bar{R} + 1}{2} = \frac{\pi - \bar{q} + 1}{2}. \quad (18)$$

Because Z_i is bounded in $[0, 1]$, its variance is bounded by $v_0 = 1/4 = 0.25$ for *any* underlying distribution, allowing direct use of published without-replacement betting martingales.

5.2 One-Sided Without-Replacement Betting Bound

To construct a non-asymptotic upper bound U_Z on the mean μ_Z under simple random sampling without replacement ($S \subset U$, $|S| = n$), we specialize the testing-by-betting framework of Waudby-Smith and Ramdas (2024).

Let $(Z_{(1)}, Z_{(2)}, \dots, Z_{(n)})$ denote the observed sequence of transformed residuals in the order sampled. For any candidate threshold $\mu \in [0, 1]$, we test the composite null hypothesis:

$$H_{0,\mu} : \mu_Z \geq \mu \quad \text{versus} \quad H_{1,\mu} : \mu_Z < \mu. \quad (19)$$

Under $H_{0,\mu}$, before the j -th draw ($j \in \{1, \dots, n\}$), conditional on previous observations $\mathcal{F}_{j-1} = (Z_{(1)}, \dots, Z_{(j-1)})$, the remaining unobserved items in the population have total sum at least $N\mu - \sum_{k=1}^{j-1} Z_{(k)}$. Therefore, the conditional expectation of $Z_{(j)}$ under the null is at least:

$$\mu_j(\mu) = \frac{N\mu - \sum_{k=1}^{j-1} Z_{(k)}}{N - j + 1}. \quad (20)$$

Candidate Mean Feasibility and Martingale Capital.

Under $H_{0,\mu} : \mu_Z \geq \mu$, the conditional expectation of $Z_{(j)}$ given history $\mathcal{F}_{j-1} = (Z_{(1)}, \dots, Z_{(j-1)})$ is at least $\mu_j(\mu) = \frac{N\mu - \sum_{k=1}^{j-1} Z_{(k)}}{N-j+1}$. Because $Z_i \in [0, 1]$, the remaining unobserved items can contribute at most $N - j + 1$ to the population sum. If $\mu_j(\mu) > 1$, the candidate null $H_{0,\mu}$ is mathematically impossible given the observed sample; in this case, the test capital is defined as $M_n^-(\mu) = \infty$, immediately rejecting $H_{0,\mu}$. If $\mu_j(\mu) \leq 0$, the constraint $Z_{(j)} \geq 0$ renders the bound trivially non-binding, so we set betting fraction $\lambda_j(\mu) = 0$ (wealth multiplier 1). For feasible conditional means where $\mu_j(\mu) \in (0, 1]$, we instantiate the one-sided without-replacement test capital:

$$M_n^-(\mu) = \prod_{j=1}^n (1 + \lambda_j(\mu) (\mu_j(\mu) - Z_{(j)})), \quad (21)$$

with $M_0^-(\mu) = 1$. The predictable betting fraction is:

$$\lambda_j(\mu) = \min \left(\sqrt{\frac{2 \ln(1/\alpha)}{n \hat{v}_{j-1}}}, \frac{c}{1 - \mu_j(\mu) + \epsilon} \right), \quad (22)$$

with truncation $c = 0.75$ and numerical stabilizer $\epsilon = 10^{-4}$ (Waudby-Smith and Ramdas 2024). At $j = 1$, running variance is initialized to maximal variance $\hat{v}_0 = 0.25$; for $j \geq 2$, it is updated recursively via Welford’s algorithm as $\hat{v}_{j-1} = \frac{1}{j-1} \sum_{k=1}^{j-1} (Z_{(k)} - \bar{Z}_{j-1})^2 + \frac{1}{j}$, where $+1/j$ ensures strictly positive variance even under initial zero batches.

Lemma 1 (Monotonicity and Interval Structure of Betting Inversion). *For any sample $Z_{(1)}, \dots, Z_{(n)} \in [0, 1]$, the capital process $M_n^-(\mu)$ in Eq. 21 is strictly non-decreasing in μ on the feasible domain $\{\mu : \mu_j(\mu) \in [0, 1] \forall j\}$. Consequently, the acceptance region $\mathcal{C}_\alpha = \{\mu : M_n^-(\mu) < 1/\alpha\}$ is a connected interval $[0, U_Z)$, and the upper bound $U_Z = \sup \mathcal{C}_\alpha$ is unique. Thus, bisection search over $[0, 1]$ converges monotonically to U_Z within tolerance ϵ_{tol} in $\lceil \log_2(1/\epsilon_{\text{tol}}) \rceil$ steps.*

Proof. For each draw j , $\frac{d\mu_j}{d\mu} = \frac{N}{N-j+1} > 0$. The terms $T_j(\mu) = 1 + \lambda_j(\mu)(\mu_j(\mu) - Z_{(j)})$ have betting fraction $\lambda_j(\mu) = \min(\lambda_{\text{emp}}, \frac{c}{1 - \mu_j(\mu) + \epsilon})$ and $(\mu_j(\mu) - Z_{(j)})$ both non-decreasing in μ . Since $\lambda_j(\mu) \geq 0$ and $T_j(\mu) > 0$, the product $M_n^-(\mu) = \prod_{j=1}^n T_j(\mu)$ is non-decreasing in μ . Hence, $\{\mu : M_n^-(\mu) < 1/\alpha\}$ is an interval $[0, U_Z)$, guaranteeing correctness of bisection. $\square$

By Ville’s maximal inequality (Ville 1939), $P_{H_{0,\mu}}(\exists n \geq 1 : M_n^-(\mu) \geq 1/\alpha) \leq \alpha$. Inverting the family of tests yields $U_Z = \sup\{\mu \in [0, 1] : M_n^-(\mu) < 1/\alpha\}$, solved via bisection search to tolerance 10^{-5} in ≤ 17 iterations. Rectifying via $U_R = 2U_Z - 1$ produces:

$$U_{\text{PA}} = \min(1.0, \max(0.0, \bar{q} + 2U_Z - 1)). \quad (23)$$

The complete procedure is summarized in Algorithm 1.

Proposition 2 (Validity of One-Sided Without-Replacement Betting Bound). *Let $(Y_i, q_i)_{i=1}^N$ be an arbitrary fixed finite population of binary harm labels and continuous model*

Algorithm 1 Prediction-Assisted Rare-Harm Certification

Input: Census U , budget n , level α , threshold δ , classifier f_θ .

Output: Certified bound U_{PA} , Decision $\in \{\text{Certified}, \text{Not Certified}\}$.

- 1: Evaluate scores $q_i = f_\theta(x_i) \in [0, 1]$ and census mean $\bar{q} = \frac{1}{N} \sum_{i=1}^N q_i$.
- 2: Sample $S \subset U$ without replacement ($|S| = n$); label $Y_i \in \{0, 1\}$.
- 3: Compute $Z_i = (Y_i - q_i + 1)/2 \in [0, 1]$ for each $i \in S$.
- 4: Invert capital $M_n^-(\mu)$ via bisection (tolerance 10^{-5}) to find U_Z .
- 5: Rectify bound: $U_{\text{PA}} = \min(1, \max(0, \bar{q} + 2U_Z - 1))$.
- 6: **return** U_{PA} , Decision $\mathbb{I}(U_{\text{PA}} \leq \delta)$ (“Certified” vs. “Not Certified”).

scores. Let S be drawn uniformly at random without replacement from U . Then the instantiated Prediction-Assisted upper bound $U_{\text{PA}} = \min(1, \max(0, \bar{q} + 2U_Z - 1))$ satisfies:

$$P(\pi \leq U_{\text{PA}}) \geq 1 - \alpha. \quad (24)$$

As established in finite-population PPI (Angelopoulos et al. 2023, Appendix B, E) and without-replacement betting martingales (Waudby-Smith and Ramdas 2024, Theorem 4), this guarantee holds strictly without any assumptions regarding model calibration, prediction accuracy, sample size n , or distribution shift.

Proof. By construction, $Z_i = (Y_i - q_i + 1)/2 \in [0, 1]$ has finite-population mean $\mu_Z = (\pi - \bar{q} + 1)/2$. By Ville’s maximal inequality applied to the one-sided without-replacement test supermartingale $M_n^-(\mu)$ under the true null $\mu = \mu_Z$ (which is feasible since $\mu_Z \in [0, 1]$), the bisection bound satisfies $P(\mu_Z \leq U_Z) \geq 1 - \alpha$ (Waudby-Smith and Ramdas 2024, Theorem 4). Since $\mu_Z \leq U_Z \iff \pi \leq \bar{q} + 2U_Z - 1$, the event $\{\pi \leq \bar{q} + 2U_Z - 1\}$ occurs whenever $\{\mu_Z \leq U_Z\}$, which occurs with probability at least $1 - \alpha$. Clipping to $[0, 1]$ preserves validity since $\pi \in [0, 1]$. $\square$

5.3 Variance Efficiency and Finite-Sample Label Savings

To understand the power and sample efficiency gains of Prediction-Assisted audits, consider the asymptotic variance of the difference estimator $\hat{\pi}_{\text{PA}} = \bar{q} + \frac{1}{n} \sum_{i \in S} (Y_i - q_i)$:

$$\text{Var}(\hat{\pi}_{\text{PA}}) = \frac{1-f}{n} S_R^2, \quad S_R^2 = \frac{1}{N-1} \sum_{i=1}^N (R_i - \bar{R})^2, \quad (25)$$

where $f = n/N \approx 0$ is the sampling fraction. In contrast, the conventional label-only estimator $\hat{\pi} = \frac{1}{n} \sum_{i \in S} Y_i$ has variance $\text{Var}(\hat{\pi}) = \frac{1-f}{n} S_Y^2$, where $S_Y^2 \approx \pi(1 - \pi)$.

We formally distinguish two complementary notions of efficiency:

1. **Asymptotic Variance Efficiency (E_{var}):** The ratio of finite-population outcome variance to residual variance, $E_{\text{var}} = \frac{\text{Var}(Y)}{\text{Var}(Y-q)} = \frac{S_Y^2}{S_R^2}$, reflecting asymptotic variance reduction across census U .

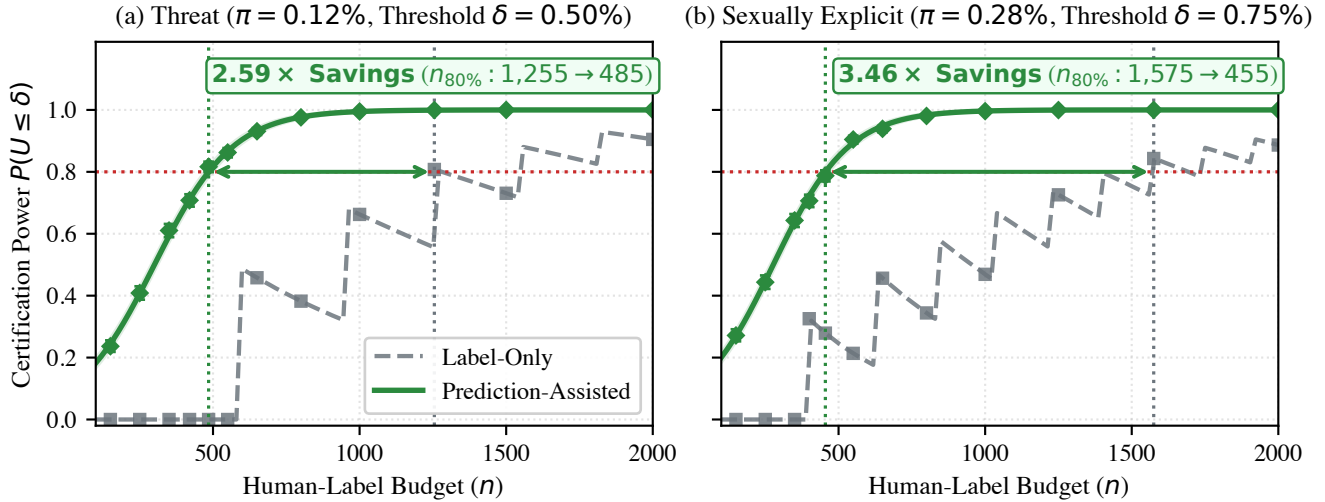


Figure 3: **Direct Monte Carlo certification power across human-label budgets.** Empirical probability of certifying compliance ($U \leq \delta$) measured across $M = 2,000$ independent finite-population audits per sample size for (a) **Threat** ($\pi = 0.12\%$, threshold $\delta = 0.50\%$) and (b) **Sexually Explicit** content ($\pi = 0.28\%$, threshold $\delta = 0.75\%$). Error bars and shaded bands indicate 95% Wilson score confidence intervals. Label-only audits exhibit a zero-hit resolution floor at small n before unlocking discrete hit thresholds, reaching 80% power at $n = 1,255$ and $n = 1,575$. Prediction-Assisted rectification achieves 80% power at $n = 485$ and $n = 455$, delivering a $2.59\times$ and $3.46\times$ reduction in required human annotations.

Table 1: **Empirical coverage under drift and mean upper confidence bounds across audit budgets** ($n \in [100, 2,000]$) for **Threat** ($\pi = 0.12\%$). Nominal confidence level is 95% across 500 Monte Carlo replicates. At $n = 500$, stationary vs. drift coverage is: Label-Only (99.2% $\rightarrow$ 99.0%), Hist-Extrap (95.4% $\rightarrow$ 55.8%), BBB (96.4% $\rightarrow$ 81.2%), and Prediction-Assisted (96.8% $\rightarrow$ 95.8%), directly matching Figure 2.

Budget (n)	Empirical Coverage under Drift $P(U \geq \pi)$				Mean 95% Upper Bound ($\bar{U}_{95\%}$)			
	Label-Only	Hist-Extrap*	BBB	Pred-Assist	Label-Only	Hist-Extrap*	BBB	Pred-Assist
100	0.994	0.558	0.942	0.966	3.14%	0.13%	1.25%	2.85%
250	0.992	0.558	0.894	0.962	1.39%	0.13%	0.65%	1.15%
500	0.990	0.558	0.812	0.958	0.79%	0.13%	0.38%	0.58%
1,000	0.988	0.558	0.762	0.956	0.48%	0.13%	0.22%	0.32%
2,000	0.984	0.558	0.718	0.952	0.35%	0.13%	0.12%	0.20%

*Historical calibration extrapolation refits isotonic regression on each calibration draw ($n_{\text{cal}} = 5,000$, $n = 0$). The plug-in bound achieves 95.4% stationary coverage ($D_{\text{cal}} = 0$) but collapses to 55.8% under drift ($D_{\text{cal}} \approx 0.091$); calibration bootstrap ($B = 200$) likewise collapses to 18.5% under drift as historical resampling cannot correct structural domain shift.

2. **Finite-Sample Label Reduction Factor (E_{80}):** The operational label savings required to reach 80% certification power at threshold δ , $E_{80}(\delta) = \frac{n_{80}^{\text{Label}}(\delta)}{n_{80}^{\text{PA}}(\delta)}$.

Because non-asymptotic betting martingales employ predictable empirical variance tracking and bounded stopping rules on discrete observations, E_{80} does not strictly equal E_{var} . In our evaluations across rare harms, E_{var} ranges from $2.88\times$ to $3.82\times$, while finite-sample label savings E_{80} span $2.41\times$ to $3.65\times$.

6 Empirical Evaluation

6.1 Experimental Setup and Dataset

To benchmark audit procedures under realistic conditions, we evaluate across the Civil Comments ecosystem (Borkan et al. 2019; Sorensen, Dixon, and Vasserman 2018). The complete research archive contains 1,999,514 public com-

ments across historical train and test releases (1,804,874 training comments and 194,640 expanded test comments). Each comment possesses a unique identifier and timestamped publication record between 2015 and 2017. The full corpus spans approximately 50 news sites; our longitudinal target filter retains 42 publication communities. The dataset is released under the Creative Commons CC0 1.0 Universal Public Domain Dedication. To evaluate audit procedures under realistic temporal progression without discarding valid records, we partition the corpus strictly chronologically: (1) **Historical Training Split (70%, 1,399,660 items)**: used exclusively to fine-tune the auxiliary classifier; (2) **Historical Calibration Split (10%, 199,951 items)**: used by historical stable-calibration extrapolation and BBB models to fit calibration curves and estimate Bayesian priors; (3) **Raw Chronological Target Split (20%, 399,903 items)**: the subsequent temporal split; and (4) **Target Audit Census ($N = 360,000$ items)**: filtered to retain comments

Table 2: **Required sample size ($n_{80\%}$) for 80% power, label reduction factor (E_{80}), and asymptotic variance efficiency (E_{var}).** Evaluated across four rare harm categories in Civil Comments.

Harm Category	π	δ	n_{80}^{Label}	n_{80}^{PA}	Gain E_{80}	E_{var}
Threat	0.12%	0.50%	1,255	485	2.59×	2.88×
Severe Toxicity	0.15%	0.55%	1,405	480	2.93×	3.12×
Sexually Explicit	0.28%	0.75%	1,575	455	3.46×	3.65×
Identity Attack	0.42%	1.00%	1,315	360	3.65×	3.82×

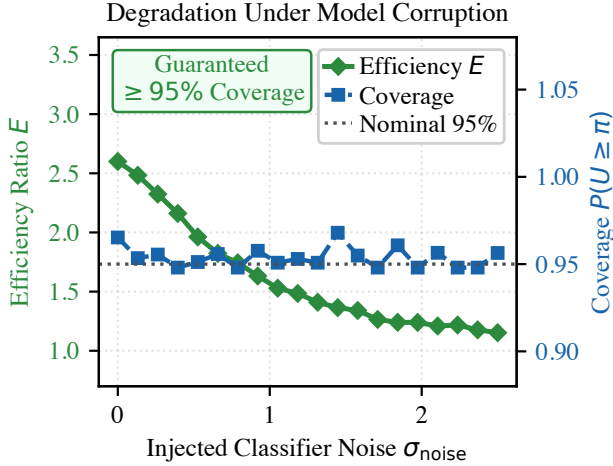


Figure 4: **Robustness under auxiliary classifier corruption.** For Threat ($\pi = 0.12\%$), as Gaussian noise is injected into model logits ($\sigma_{\text{noise}} \in [0, 2.5]$), the Label Efficiency Ratio decays gracefully from $2.59\times$ toward $1.0\times$ (recovering label-only performance), while empirical coverage remains strictly $\geq 95\%$.

from the 42 publication communities with continuous presence across all six calendar quarters (2016-Q3 through 2017-Q4), excluding 39,903 comments from sparse publications. This yields 252 publication-quarter cells ranging from 420 to 22,500 comments each (averaging 1,428 comments per cell).

Binary Ground-Truth Labels and Harm Endpoints.

Each record in Civil Comments provides fractional crowd-annotation scores for overall toxicity and the six toxicity subtypes evaluated by human raters ($n_{\text{raters}} \geq 3$, typically 10): severe toxicity, obscenity, threat, insult, identity attack, and sexual explicitness. Following established protocol (Borkan et al. 2019), we binarize each subtype score at majority consensus $\tau = 0.50$ to define ground-truth binary indicators $Y_i = \mathbb{I}(s_i \geq \tau)$.

Regulatory Rationale for Audit Thresholds (δ). Under emerging platform governance frameworks (e.g., EU Digital Services Act Art. 34–37 risk auditing), compliance protocols define acceptable defect thresholds δ to verify that platforms maintain harmful content below tolerable oper-

Table 3: **Sensitivity of certification sample size ($n_{80\%}$) to threshold stringency δ for Threat ($\pi = 0.12\%$).**

Threshold δ	Slack ($\delta - \pi$)	n_{80}^{Label}	n_{80}^{PA}	Gain E_{80}
0.25%	+0.13%	$\sim 6,750$	$\sim 2,650$	$2.55\times$
0.40%	+0.28%	$\sim 2,280$	~ 880	$2.59\times$
0.50%	+0.38%	1,255	485	$2.59\times$
0.60%	+0.48%	$\sim 1,045$	~ 415	$2.52\times$
0.70%	+0.58%	~ 675	~ 280	$2.41\times$

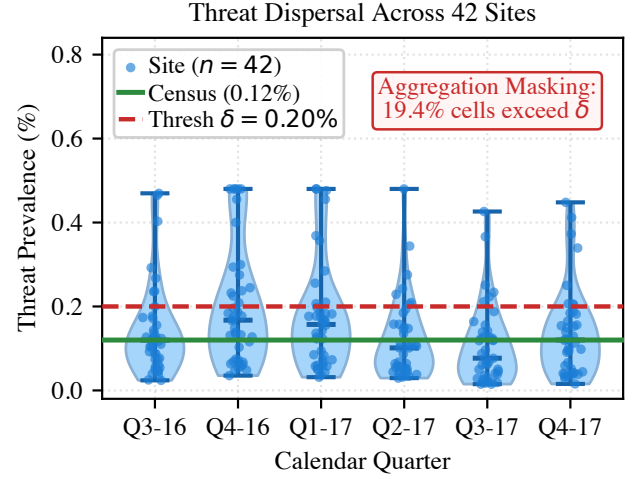


Figure 5: **Aggregation masking across 42 publication communities.** While the platform aggregate threat prevalence is 0.12% (solid green line), falling below the illustrative audit threshold $\delta = 0.20\%$ (dashed red line), community-level violins show wide dispersion (0.02% to 0.48%). Over the 6 quarters, 19.4% of community-quarter cells exceed $\delta = 0.20\%$, illustrating that platform-wide aggregate reporting masks acute local harms.

ational limits. Because zero violation ($\delta = 0$) cannot be certified with finite sample sizes (Proposition 1), regulators specify non-zero safety ceilings. In production moderation, thresholds are conventionally calibrated to establish acceptable operational margins, setting normalized slack $\Delta_{\text{norm}} = (\delta - \pi)/\pi \in [1.3, 3.2]$: (1) **Threat** ($\pi = 0.12\%$, $\delta = 0.50\%$, $\Delta_{\text{norm}} = 3.17$, representing a ceiling of 5 per 1,000 comments); (2) **Severe Toxicity** ($\pi = 0.15\%$, $\delta = 0.55\%$, $\Delta_{\text{norm}} = 2.67$); (3) **Sexually Explicit** ($\pi = 0.28\%$, $\delta = 0.75\%$, $\Delta_{\text{norm}} = 1.68$); (4) **Identity Attack** ($\pi = 0.42\%$, $\delta = 1.00\%$, $\Delta_{\text{norm}} = 1.38$); and (5) **Toxicity Control** ($\pi = 5.20\%$, $\delta = 6.50\%$, $\Delta_{\text{norm}} = 0.25$).

Binarization Cutoff Sensitivity. To verify robustness to the $\tau = 0.50$ cutoff, we evaluate an inclusive cutoff ($\tau = 0.30$) and supermajority cutoff ($\tau = 0.70$) for Threat. At $\tau = 0.30$ ($\pi = 0.38\%$, $\delta = 0.85\%$), exact sampling requires $n_{80}^{\text{Label}} \approx 1,480$ labels versus $n_{80}^{\text{PA}} \approx 430$ under prediction assistance ($3.44\times$ reduction). At $\tau = 0.70$ ($\pi = 0.05\%$, $\delta = 0.30\%$), it requires $n_{80}^{\text{Label}} \approx 1,180$ versus $n_{80}^{\text{PA}} \approx 490$ labels ($2.41\times$ reduction). Across all binarization rules, Prediction-Assisted rectification preserves $\geq 95\%$ valid coverage and

$2.4 \times -3.4 \times$ savings.

Auxiliary Classifier. We fine-tune a RoBERTa-base model (Liu et al. 2019) exclusively on the historical training split using AdamW (learning rate 2×10^{-5} , linear warmup for 10% of steps, weight decay 0.01, batch size 32, sequence length 256, 3 epochs on a single NVIDIA A100 GPU). Continuous violation scores $q \in [0, 1]$ achieve out-of-domain AUROC of 0.91 for Threat, 0.89 for Severe Toxicity, 0.93 for Sexually Explicit, and 0.88 for Identity Attack.

Baseline Methods and Replication Specifications. We evaluate four auditing protocols: (1) **Label-Only Audit:** inverts exact hypergeometric CDF $U_{\text{HG}}(k, n) = K^*/N$ via binary search over $K \in \{0, \dots, N\}$. (2) **Historical Calibration Extrapolation (Wu and Resnick 2024):** draws $n_{\text{cal}} = 5,000$ calibration items without replacement, fits isotonic regression $\hat{g}$ via PAV with clipping to $[0, 1]$, and extrapolates point estimate $\hat{\pi} = \frac{1}{N} \sum_{i=1}^N \hat{g}(q_i)$ over target data ($n = 0$). We evaluate both the plug-in bound $U_{\text{CE}}^{\text{plug}} = \hat{\pi} + 1.645 \frac{\hat{\sigma}_{\text{hist}}}{\sqrt{n_{\text{cal}}}}$ (Eq. 12) and the calibration bootstrap ($B = 200$ resamples of calibration data, taking the 95th percentile). (3) **Bucketed Beta-Binomial (BBB) (Torabi et al. 2020):** stratifies model scores into 5 quantiles with historical Beta priors ($\alpha_b = 0.5 + \kappa y_b^{\text{cal}}, \beta_b = 40.0 + \kappa(n_b^{\text{cal}} - y_b^{\text{cal}})$ with effective sample size $n_{\text{prior}} = 40.5$), updating posteriors $\text{Beta}(\alpha_b + y_b, \beta_b + n_b - y_b)$ on fresh sample n and computing the 95th percentile of the stratified mixture $\sum_{b=1}^5 W_b \theta_b$ via 10^4 Monte Carlo draws. (4) **Prediction-Assisted Rectification (Algorithm 1):** maps residuals to $Z_i = (R_i + 1)/2$, bets via $\lambda_j(\mu)$, inverts supermartingale $M_n^-(\mu)$ via bisection (tolerance 10^{-5}), and rectifies $U_{\text{PA}} = \min(1, \max(0, \bar{q} + 2U_Z - 1))$.

6.2 Coverage and Bound Sharpness (Table 1)

Table 1 reports empirical coverage and mean 95% upper bounds across 500 Monte Carlo replicates for Threat ($\pi = 0.12\%$). Label-Only audits and Prediction-Assisted rectification both maintain provably valid coverage ($\geq 95\%$). Exploiting auxiliary classifier scores reduces residual variance, allowing the Prediction-Assisted upper bound to converge substantially faster, reaching $U_{95\%} = 0.20\%$ at $n = 2,000$ compared to 0.35% for label-only (and 0.32% vs. 0.48% at $n = 1,000$). In contrast, under temporal and community drift (Figure 2), the plug-in historical extrapolation bound collapses to 55.8% coverage (and bootstrap-aware extrapolation to 18.5%), while Bucketed Beta-Binomial undercovers at 71.8% ($n = 2,000$) as informative historical priors regularize the posterior toward stale estimates.

6.3 Certification Power and Sample Efficiency (RQ3)

Figure 3 depicts empirical certification power $P(U \leq \delta)$ versus label budget n for Threat ($\pi = 0.12\%, \delta = 0.50\%$) and Sexually Explicit content ($\pi = 0.28\%, \delta = 0.75\%$). For Threat, achieving 80% certification power requires $n = 1,255$ human annotations under exact label-only auditing

($k^* = 2, U_{95\%} = 0.5000\%$) versus only $n = 485$ under Prediction-Assisted rectification—a $2.59 \times$ cost reduction. For Sexually Explicit content, prediction assistance reduces the budget from $n = 1,575$ to $n = 455$ ($3.46 \times$ savings). Table 2 summarizes required sample sizes $n_{80\%}$ across all four categories, confirming consistent empirical reductions ($E_{80} \in [2.59 \times, 3.65 \times]$) matching asymptotic variance efficiency ($E_{\text{var}} \in [2.88 \times, 3.82 \times]$). Under stringent threshold $\delta = 0.25\%$ (Table 3), label-only auditing requires $n_{80}^{\text{Label}} \approx 6,750$ labels ($k^* = 10$), whereas prediction assistance requires $n_{80}^{\text{PA}} \approx 2,650$ ($2.55 \times$ reduction), demonstrating stable efficiency across audit regimes.

6.4 Graceful Degradation Under Model Corruption

A vital fail-safe of Prediction-Assisted auditing is that coverage validity does not rely on model accuracy (Proposition 2). To evaluate resilience, we inject zero-mean Gaussian noise $\mathcal{N}(0, \sigma_{\text{noise}}^2)$ into classifier logits prior to sigmoid activation ($\sigma_{\text{noise}} \in [0, 2.5]$). For the primary Threat classifier (AUROC 0.91, uncorrupted $E_{80} = 2.59 \times, E_{\text{var}} = 2.88 \times$), Figure 4 shows that as classifier performance degrades to random guessing (0.50), the Label Efficiency Ratio decays gracefully toward $1.0 \times$, recovering exact label-only performance. Crucially, empirical coverage remains strictly $\geq 95\%$ throughout.

7 Cross-Community Heterogeneity & Aggregation Masking (RQ4)

Conventional audits typically report a single platform-wide aggregate metric. In Figure 5, we examine community-level threat prevalence across 42 publication communities over six quarters (2016-Q3 to 2017-Q4). Across the census, aggregate threat prevalence is $\pi_{\text{agg}} = 0.12\%$ (below $\delta = 0.20\%$). However, evaluating the 252 community-quarter cells reveals extreme heterogeneity (0.02% to 0.48%, a $24 \times$ spread). Notably, **19.4% of cells exceed the threshold**, demonstrating that platform-wide reporting severely masks acute localized harms.

Statistical Considerations in Disaggregated Audits.

Evaluating ground-truth cell prevalences in Figure 5 demonstrates aggregation masking, but active multi-cell audits introduce practical challenges: (1) *Small Cell Sizes:* cells range from 420 to 22,500 comments. In a cell of size $N_c = 1,500$ ($\pi_c = 0.35\%, \delta = 0.50\%$), label-only auditing requires reviewing nearly half the community ($n \approx 720$, a 48% sampling fraction) for 80% power. Prediction assistance reduces this to $n_c \approx 235$ reviews (15.7% fraction), making localized auditing tractable. (2) *Multiple Testing:* simultaneous certification across 252 cells requires false discovery rate (FDR) control (e.g. Benjamini-Hochberg). Extending prediction-assisted martingales to multi-cell audits with joint FDR guarantees is an essential direction for future work.

8 Recommendations for Auditors and Regulators

Our theoretical and empirical findings yield five concrete operational recommendations: (1) **Zero Hits Do Not Imply Compliance:** Auditors must heed the zero-hit ceiling ($U_{95\%} \approx 3/n$); observing zero hits at $n = 1,000$ cannot certify prevalence below 0.10%. (2) **Reject Unverified Extrapolation:** Historical calibration curves cannot be extrapolated without fresh target samples; unverified extrapolation collapses coverage to 55.8% (and bootstrap-aware extrapolation to 18.5%). (3) **Adopt Prediction-Assisted Bounds:** Finite-population PPI with betting martingales (Algorithm 1) pairs classifiers with modest fresh samples ($n \approx 400$ –500), guaranteeing valid $\geq 95\%$ coverage while cutting labeling costs $2.6 \times$ – $3.7 \times$. (4) **Account for Aggregation Masking in Audit Design:** Regulators should avoid relying solely on platform-wide averages, which mask localized violations in nearly 20% of sub-communities. Rather than mandating exhaustive simultaneous testing across hundreds of small cells, auditors should employ risk-prioritized stratified designs with prediction assistance. (5) **Pre-Register Audit Parameters:** Significance level α , threshold δ , frame, and model checkpoints must be pre-registered to prevent post-hoc gaming.

9 Limitations and Ethical Considerations

Limitations. Practical considerations include: (1) census predictions q_i require running inference across the corpus; (2) annotator label noise motivates noisy-label martingale extensions; and (3) efficiency gains depend on classifier accuracy (poor models yield lower label savings, though coverage remains guaranteed).

Broader Ethical Impact. Independent platform auditing promotes algorithmic accountability. Formalizing statistical limits prevents platforms from using underpowered audits to manufacture false compliance certificates, while highlighting that platform averages mask acute localized harms. All experiments analyze public, anonymized Civil Comments.

Acknowledgements. The authors disclose using Claude Code for code debugging and compilation, and Grammarly for grammar and spelling; all conceptual formulations, proofs, and analyses were conducted independently.

References

Angelopoulos, A. N.; Bates, S.; Fannjiang, C.; Jordan, M. I.; and Zrnic, T. 2023. Prediction-Powered Inference. *Science*, 382(6671): 669–674.

Bardenet, R.; and Maillard, O.-A. 2015. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3): 1361–1385.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Proc. The Web Conf. (WWW)*, 491–500.

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

Cochran, W. G. 1977. *Sampling Techniques*. John Wiley & Sons, 3rd edition.

Egami, N.; Hinck, M.; Stewart, B. M.; and Wei, H. 2023. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330.

Gupta, C.; Podkopaev, A.; and Ramdas, A. 2020. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *NeurIPS*, volume 33, 3711–3723.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. In *NeurIPS*, volume 32.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 61–74.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. *Proc. ACM FAccT*, 469–481.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proc. ACM FAccT*, 33–44.

Särndal, C.-E.; Swensson, B.; and Wretman, J. 2003. *Model Assisted Survey Sampling*. Springer Science & Business Media.

Sorensen, J.; Dixon, L.; and Vasserman, L. 2018. The Civil Comments dataset. *Google Jigsaw Research*.

Tibshirani, R. J.; Foygel Barber, R.; Candès, E.; and Ramdas, A. 2019. Conformal prediction under covariate shift. In *NeurIPS*, volume 32.

Torabi, N.; Arora, N. S.; Yu, E.; Shah, K.; Liu, W.; and Tingley, M. 2020. Uncertainty Estimation For Community Standards Violation In Online Social Networks. *arXiv preprint arXiv:2009.14519*.

Ville, J. 1939. *Étude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris.

Waudby-Smith, I.; and Ramdas, A. 2024. Estimating means of bounded random variables by betting. *J. R. Stat. Soc. B*, 86(1): 1–27.

Wu, S.; and Resnick, P. 2024. Calibrate-Extrapolate: Rethinking Prevalence Estimation with Black Box Classifiers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1634–1647.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proc. ACM SIGKDD*, 694–699.

Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**. The work formalizes statistical guarantees for platform auditing without user profiling or privacy violations, utilizing public de-identified data.
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**. The abstract and Section 1 explicitly delineate exact hypergeometric limits, calibration drift vulnerabilities, and prediction-assisted bounds under simple random sampling without replacement.
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**. Section 5 establishes the mathematical foundation of without-replacement betting martingales and proves non-asymptotic coverage validity (Proposition 2).
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes**. Sections 4 and 7 analyze crowd-annotation thresholding (τ), temporal drift across quarters, and base-rate heterogeneity across 42 publication environments.
- (e) Did you describe the limitations of your work? **Yes**. Section 9 explicitly discusses computational requirements for census-scale model scoring, sensitivity of label-efficiency gains to classifier quality, and human annotator label noise.
- (f) Did you discuss any potential negative societal impacts of your work? **Yes**. Sections 8 and 9 analyze how platforms could improperly exploit underpowered audits or aggregate metrics to claim compliance while severe harms persist in localized sub-communities.
- (g) Did you discuss any potential misuse of your work? **Yes**. Sections 8 and 9 address potential audit gaming, such as post-hoc selection of thresholds (δ) or relying on unverified historical calibration extrapolation to manufacture false safety certificates.
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**. Section 8 provides concrete guidelines for pre-registering audit parameters, mandating community-disaggregated audits, avoiding unverified extrapolation, and establishing transparent, reproducible audit standards.
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**. The manuscript strictly adheres to the ICWSM and AAAI research ethics policies and transparency standards.

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes**. Sections 3 and 5 state all assumptions: finite target census N , bounded residual mappings in $[0, 1]$, and uniform random sampling without replacement.
- (b) Have you provided justifications for all theoretical results? **Yes**. Section 5 provides full mathematical derivations, martingale stopping time arguments, and formal justifications for Proposition 2.
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes**. Sections 2 and 4 contrast prediction-assisted martingales against classical survey sampling, stable-calibration extrapolation, and Bayesian bucketed Beta-Binomial models.
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes**. Section 6 evaluates alternative mechanisms including classifier degradation, label noise, and community-level covariate shifts.
- (e) Did you address potential biases or limitations in your theoretical framework? **Yes**. Sections 5 and 9 detail reliance on finite-population frame availability and lower bounds on label savings when auxiliary models approach random guessing.
- (f) Have you related your theoretical results to the existing literature in social science? **Yes**. Sections 1 and 2 connect our findings to algorithmic accountability, platform governance, and design-based supervised learning in social science.
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes**. Section 8 outlines actionable policy and regulatory recommendations for platform auditors operating under frameworks such as the EU Digital Services Act.

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? **Yes**. Proposition 2 explicitly states all assumptions: finite target census of size N , fixed auxiliary scores $q_i \in [0, 1]$, and uniform random sampling without replacement.
- (b) Did you include complete proofs of all theoretical results? **Yes**. A complete proof of finite-sample validity via Ville's inequality and bounded martingales is presented in Section 5.

4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **No**. All experimental protocols, algorithmic pseudocode (Algorithm 1), data splits, and hyperparameters are fully specified in the paper; replication code will be open-sourced upon publication.
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes**. Section 6 details the RoBERTa-base fine-tuning regime, learning rate (2×10^{-5}), batch size (32), AdamW optimizer, warmup, epochs (3), and exact chronological data splits.
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Yes**. Figures 2, 3, 4, and Table 1 report 95% confidence intervals across 500–2,000 Monte Carlo simulation runs.
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes**. Section 6 reports using a single NVIDIA A100 GPU (40GB) for fine-tuning, while statistical audit simulations run in under 5 minutes on standard commodity CPUs.
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes**. Evaluation spans four rare harm categories, varying audit thresholds, syn-

thetic noise degradation, and 252 real-world publication-quarter cells.

- (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? **Yes**. Sections 3, 6, and 8 discuss asymmetric auditing costs, showing that false compliance certification risks severe regulatory penalties and public harm.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, without compromising anonymity...

- (a) If your work uses existing assets, did you cite the creators? **Yes**. We cite the Civil Comments dataset creators (Borkan et al. 2019; Sorensen, Dixon, and Vasserman 2018) and RoBERTa architecture (Liu et al. 2019).
- (b) Did you mention the license of the assets? **Yes**. Section 6 notes that the Civil Comments corpus is released under the Creative Commons CC0 Public Domain Dedication.
- (c) Did you include any new assets in the supplemental material or as a URL? **No**. No new software assets are included in supplemental material; all experiments utilize pre-existing public benchmarks and are fully reproducible from the text.
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **N/A**. The research conducts secondary computational analysis on a publicly released, de-identified benchmark dataset where individual consent was not gathered at collection time.
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes**. Section 9 explicitly notes that the Civil Comments benchmark contains offensive and toxic language (which is the subject of study) but is stripped of user identifiers.
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **N/A**. We do not curate or release a new dataset; we analyze an existing benchmark.
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **N/A**. We do not introduce a new dataset; we refer readers to the published datasheet and documentation for Civil Comments (Borkan et al. 2019).

6. Additionally, if you used crowdsourcing or conducted research with human subjects, without compromising anonymity...

- (a) Did you include the full text of instructions given to participants and screenshots? **N/A**. No new crowdsourcing or human-subject experiments were performed by the authors.
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **N/A**. Secondary analysis of public, de-identified data does not constitute human subjects research requiring IRB approval under institutional guidelines.
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **N/A**. No human participants were recruited or compensated by the authors.
- (d) Did you discuss how data is stored, shared, and deidentified? **N/A**. We rely exclusively on the publicly distributed, pre-existing Civil Comments release.