ISSN: 3143-8458

Wissen PUBLICATION GROUP

Journal of
Artificial Intelligence and Data Analytics

Research Article

Volume: 1 | Issue: 3 | November 2026 | Pages 01-07

The Prevention and Observability Frontier in AI Action Auditing

Kwan Hong Tan

Singapore University of Social Sciences, Singapore

*Corresponding Author: Kwan Hong Tan, Singapore University of Social Sciences, Singapore. E-mail: khtan055@suss.edu.sg

Submitted: 16 October 2026 **Accepted:** 29 October 2026 **Published:** 03 November 2026

Citation: Tan, K.H. (2026). The Prevention and Observability Frontier in AI Action Auditing. Journal of Artificial Intelligence and Data Analytics, 1(3), 01-07. DOI: doi.org/10.67927/JAIDataAnalytics/2026(1)115.

Abstract

An artificial intelligence audit can both reveal a defective action and prevent that action from being released. Consequently, the audit sample changes the residual loss that it is intended to estimate. This paper develops a finite-population framework for this dual role of verification. A model-assisted estimator targets the loss remaining in the actual unaudited batch, and an exact error identity distinguishes this random target from average policy risk. The framework yields a finite-sample upper confidence bound, an impossibility result for unauditable actions, and a convex allocation rule balancing predicted prevention against worst-case estimation variance. In a transparent synthetic population of 6,000 actions, five policies receive the same expected budget of 1,200 audits. Under a concealed error-rate shift, the joint allocation reduces expected residual loss by 34.7% and estimation root mean squared error by 89.3% relative to risk-priority auditing. Under aligned scores, however, it incurs 34.0% more residual loss than risk-priority auditing, exposing the cost of observability. Twenty thousand audit replications per policy and scenario verify the analytical calculations and reveal conservative confidence bounds. The contribution is an explicit connection between corrective auditing, the statistical target, and resource allocation, rather than a new inverse-probability estimator. The results support auditable AI release decisions under stated assumptions, without claiming empirical validation on deployed systems.

Keywords: AI Auditing, Residual Risk, Model-Assisted Estimation, Randomized Verification, Audit Allocation, Finite-Population Inference.

Introduction

Consider an AI agent proposing thousands of changes to enterprise records. An independent check can identify an invalid change and stop its execution. After inspecting some proposals, the operator wants to know how much defective output remains in the proposals that will actually be released. This question differs from estimating the error rate of the original proposal pool. It also differs from estimating the average loss that would remain if the same randomized audit policy were repeated. The distinction matters because an audit is simultaneously a measurement and an intervention. The selective-labels literature establishes that observed labels can depend on earlier decisions and that evaluation on observed cases alone can be misleading [1]. Active testing deliberately acquires labels to improve evaluation efficiency, while active surrogate estimators learn models of unobserved errors [2,3]. These contributions make purposeful sampling a natural component of AI evaluation. For corrective audits, however, the target itself depends on the selected sample: inspected defects are removed from the release pool.

This paper asks how an operator should allocate audits when both preventing defects and measuring the defects left behind matter. Concentrating checks on high predicted risk can prevent many errors while leaving other regions almost unobservable. Conversely, allocating checks to improve statistical precision can consume capacity that would otherwise correct likely errors. We call the feasible trade-off between predicted residual loss and estimation uncertainty the prevention and observability frontier. Here, observability means statistical learnability of a bounded aggregate loss from logged randomized audits. It does not refer to reconstructing a dynamical system state. The central contribution is a finite-population analysis of realized residual loss under corrective auditing. It combines a classical generalized difference estimator with an exact random-target error identity, a conservative finite-sample certificate, and an allocation rule that makes the prevention and precision trade-off explicit. A fully specified synthetic experiment evaluates the mechanism, including a setting where the joint policy is inferior on prevention. No trained AI model, production dataset, or real-world intervention is represented as having been evaluated.

Related Research and the Specific Contribution

Inverse-probability estimation dates to Horvitz and Thompson [4]. Model-assisted survey sampling uses auxiliary predictions together with probability-sampled residual corrections [5]. Prediction-powered inference similarly combines model predictions and labeled observations to support valid inference [6]. Prediction-powered active testing further couples residualized evaluation with variance-oriented acquisition [7]. These are the statistical foundations and closest estimation precedents of the present estimator. Its algebra should not be interpreted as an independent invention of debiasing or model assistance. Conformal risk control calibrates a prediction procedure for an expected monotone loss under its stated

sampling conditions [8]. Recent work moves audit-backed guarantees into operational decisions. Manoharan studies random verification of screened power-system contingencies, including risk among the unverified trusted subset [9]. Hu et al. [10] study risk-certified update admission with imperfect verifiers and a finite secondary-verification budget. Both are closely related precedents. The present analysis concentrates on heterogeneous, bounded action losses, independent unequal-probability audits, and the covariance created when the same sampled units are both measured and corrected. Its specific addition is the joint treatment of sample-dependent post-correction loss and a prevention–variance allocation objective. Existing audit-based residual-risk guarantees are therefore acknowledged; no claim of the first audited risk certificate is made.

Within enterprise AI assurance, Tan develops a policy-gated runtime architecture; Tan studies verification allocation to suppress network error cascades; and Tan models verification debt in organizational automation [11-13]. These papers motivate the operational setting but do not supply the sampling results established here. The present contribution concerns the evidence available after a particular corrective audit, with no branching-process or organizational-dynamics claim. Table 1 summarizes the distinctions.

Table 1: Relationship to Established Approaches.

Approach	Main Target	Relationship to this Study
Survey estimation and model assistance	Fixed population totals	Provides the estimator and design-based reasoning
Active testing	Efficient model evaluation	Motivates purposeful label acquisition
Conformal risk control	Calibrated expected loss	Provides a distinct form of risk guarantee
Audited selective verification	Risk of screened or admitted outputs	Closest operational precedent
Present formulation	Loss in the realized unaudited batch	Couples correction, uncertainty, and audit allocation

Methodology

Finite Population and Audit Protocol

Fix a batch of N proposed AI actions before any audit selection. Each action i has potential defect loss y_i and known upper bound $h_i>0$. Its audit cost is $c_i>0$, and its pre-audit surrogate is g_i. The surrogate is clipped to the loss range. A loss may be a severity-weighted violation of a validation rule or an independently adjudicated defect. It is not automatically an estimate of a downstream causal harm. In particular, a prevented action must still have a well-defined label that an auditor can determine without executing it.

$$0\le y_i\le h_i, \qquad 0\le g_i\le h_i, \qquad A_i\sim\text{Bernoulli}(q_i). \tag{1}$$

The audit indicators A_i are mutually independent conditional on the frozen batch, surrogates, costs, and probabilities. Their probabilities satisfy $q_i\in[q_{min},1]$, where $q_{min}>0$. The audit reveals y_i exactly and removes its defect loss before release. All unaudited proposals are released in the basic model. The values y_i are fixed for design-based inference; they need not be independent or identically distributed. They may exhibit arbitrary common-model or task-level dependence before the batch is frozen.

This protocol has four substantive requirements. Audit selection must use logged randomization with known probabilities. Surrogates and probabilities must be fixed before the current batch's labels are revealed. Audited losses must be ascertainable and completely corrected. Finally, auditing one action must not change another action's potential defect loss. Dependencies caused by shared documents, upstream edits, or multi-agent cascades require a different model or a cluster-level audit unit.

Define the realized residual loss L, its design expectation R, and the expected audit cost C by

$$\begin{aligned} L(A) &= \sum_{i=1}^{N}(1-A_i)\,y_i, \qquad R(q)=\sum_{i=1}^{N}(1-q_i)\,y_i, \\ C(q) &= \sum_{i=1}^{N} c_i\,q_i. \end{aligned} \tag{2}$$

The randomness in L comes from audit selection. The phrase "realized residual" denotes the actual unaudited batch's defect total, not a future population mean. Dividing all loss quantities and confidence bounds by the fixed N gives loss per proposed action. A loss rate per released action has a different denominator and is not analyzed here.

Model-Assisted Estimation after Correction

The generalized difference estimator of the original total predicts every action's loss and uses inverse-probability weighting to correct prediction errors on audited actions. Subtracting the observed loss removed by the same audit gives

$$\hat{L}_g=\sum_{i=1}^{N} g_i+\sum_{i=1}^{N}\frac{A_i}{q_i}(y_i-g_i)-\sum_{i=1}^{N}A_i\,y_i. \tag{3}$$

An operationally equivalent form predicts loss among unaudited actions and corrects that prediction using audited residuals:

$$\hat{L}_g=\sum_{i=1}^{N}(1-A_i)\,g_i+\sum_{i=1}^{N}\frac{A_i(1-q_i)}{q_i}(y_i-g_i). \tag{4}$$

Only audited values of y_i are needed in either expression. The special case $g_i=0$ is denoted $\hat{L}_0$. Model assistance is optional for validity. Accurate surrogates can improve precision, but unweighted surrogate predictions alone do not provide a design-based guarantee.

Proposition 1: Conditional on the frozen batch and audit design, the estimator has zero expected error for the random target L and the following exact mean squared error:

$$\hat{L}_g-L=\sum_{i=1}^{N}\left(\frac{A_i}{q_i}-1\right)(y_i-g_i), \qquad E_A[\hat{L}_g-L]=0. \tag{5}$$

$$V_g(q;y)=\mathrm{E}_A\left[\left(\hat{L}_g-L\right)^2\right]=\sum_{i=1}^{N}\frac{1-q_i}{q_i}(y_i-g_i)^2. \quad (6)$$

Proof. Subtract Equation (2) from Equation (3) and collect the terms for each action to obtain Equation (5). Each summand has expectation zero. Independence of the audit indicators eliminates cross terms; $\mathrm{Var}(A_i/q_i)=(1-q_i)/q_i$ gives Equation (6). No model-correctness assumption is used. The zero-error statement is marginal over audit randomization. It does not assert unbiasedness conditional on the observed audit sample or conditional on the realized value of L.

The random target cannot be treated as fixed when estimating uncertainty. In the unassisted case, the same point estimator estimates the fixed expectation R, but its variance around R is

$$\mathrm{Var}_A\left(\hat{L}_0\right)=\sum_{i=1}^{N}\frac{(1-q_i)^3}{q_i}y_i^2. \quad (7)$$

At a common audit probability $q<1$, the variance in Equation (6), with $g=0$, is larger than Equation (7) by the factor $1/(1-q)^2$. The negative covariance between the estimated remaining loss and the actual remaining loss is essential: observing more defects increases the inverse-probability estimate while simultaneously reducing the defects left for release.

A Finite-Sample Certificate for the Unaudited Batch

For each action, let $d_i=\max\{g_i,h_i-g_i\}$. This quantity bounds the absolute surrogate error for every allowed y_i. Define

$$\overline{V}_g(q)=\sum_{i=1}^{N}d_i^2\frac{1-q_i}{q_i},\qquad M_g(q)=\max_{i:q_i<1}\left\{h_i-g_i,\frac{g_i(1-q_i)}{q_i}\right\}. \quad (8)$$

The maximum is zero if every action is audited. The variance envelope is exact as a worst case over the rectangular loss set: each loss may be selected at whichever endpoint, zero or h_i, is farther from its surrogate. It is therefore conservative when substantive information excludes many simultaneous worst-case residuals.

Proposition 2: For a pre-specified $0<\delta<1$, set $t=\log(1/\delta)$ and

$$\begin{aligned} r_\delta(q) &= \sqrt{2\overline{V}_g(q)t}+\frac{M_g(q)t}{3}, \\ U_g &= \min\left\{\sum_{i=1}^{N}(1-A_i)\,h_i,\ \max\{0,\hat{L}_g+r_\delta(q)\}\right\}. \end{aligned} \quad (9)$$

Then $\Pr_A\{L(A)\le U_g\}\ge 1-\delta$.

Proof. Write $L-\hat{L}_g=\sum_i W_i$, where $W_i=(1-A_i/q_i)(y_i-g_i)$. These independent, mean-zero variables have total variance no larger than $\overline{V}_g$. If $A_i=0$, then $W_i\le h_i-g_i$; if $A_i=1$, then $W_i\le g_i(1-q_i)/q_i$. For $q_i=1$, the variable is identically zero. The one-sided Bernstein inequality gives an upper deviation of $\sqrt{2\overline{V}_g t}+\frac{M_g t}{3}$. with failure probability at most exp(-t) [14]. Clipping below at zero cannot reduce coverage for a nonnegative target. Clipping above at the known maximum loss of the unaudited set preserves coverage because that maximum always bounds L.

The result is a repeated-randomization guarantee for a specified finite batch. It does not certify each action, guarantee a future batch, or establish that the residual risk meets a chosen operational tolerance. Such a tolerance is met on the reported confidence event only when the observed U_g is sufficiently small. A large bound is an informative failure to certify, even if the point estimate is reassuring.

When the Audit Design Makes Risk Unidentifiable

Proposition 3: Suppose an action with $h_i>0$ has $q_i=0$, and no additional information restricts its loss. No estimator based only on the audit transcript can be design-unbiased for residual loss for every admissible population.

Proof: Construct two populations identical on all observable quantities and all other losses. In the first, the never-audited action has zero loss; in the second, its loss is h_i. The audit transcripts have identical distributions, but the expected residual losses differ by h_i. Any transcript-measurable estimator has the same expectation in both populations and cannot be unbiased in both. For a never-audited subset, the same argument creates ambiguity equal to the sum of its loss bounds. Positive audit probabilities remove this support failure, but arbitrarily small probabilities can still make useful inference prohibitively imprecise. This distinction separates identifiability from practical certifiability.

Joint Allocation of Prevention and Observability

The actual losses are unavailable when audits are assigned. Let $s_i\in[0,h_i]$ be a pre-audit forecast used for prevention. It may equal g_i, as in the primary experiment, but the allocation forecast and estimation center need not be identical. We optimize the proxy $\tilde{R}(q)=\sum_i s_i(1-q_i)$ together with the worst-case variance envelope determined by g_i. For $\lambda>0$ and expected budget B, solve

$$\min_{q\in[q_{\min},1]^N}\left\{\tilde{R}(q)+\lambda\overline{V}_g(q)\right\}\quad\text{subject to}\quad\sum_{i=1}^{N}c_i\,q_i\le B. \quad (10)$$

The parameter λ has inverse-loss units, so its numerical value changes when the loss scale changes. A choice of λ should express the operator's valuation of precision, not be tuned against hidden current-batch outcomes. This objective penalizes variance rather than directly minimizing the full confidence radius, which also depends on M_g.

Proposition 4: For $\sum_i c_i q_{min}<B<\sum_i c_i$, Equation (10) is strictly convex and has a unique minimizer. For a multiplier $\mu\ge 0$ enforcing the budget, its coordinates satisfy

$$q_i^*=\begin{cases}\min\left\{1,\max\left\{q_{\min},\sqrt{\dfrac{\lambda d_i^2}{\mu c_i-s_i}}\right\}\right\}, & \mu c_i>s_i,\\ 1, & \mu c_i\le s_i.\end{cases} \quad (11)$$

Proof: Each $d_i>0$ because $h_i>0$. The second derivative of the ith objective term is $2\lambda d_i^2/q_i^3>0$. The feasible set is convex and compact, and an interior feasible point exists. The objective is decreasing in every coordinate, so the cost constraint binds. The Karush–Kuhn–Tucker conditions, as developed for convex programs by Boyd and Vandenberghe [15], are sufficient. For an interior coordinate, stationarity gives $-s_i-\lambda d_i^2/q_i^2+\mu c_i=0$. Solving and clipping to the interval gives Equation (11); a nonpositive denominator implies a decreasing coordinate objective

throughout the interval and hence its upper endpoint.

Expected cost is continuous and nonincreasing in μ, allowing bisection. Each iteration costs $O(N)$ operations. The boundary case $\lambda=0$ gives risk-priority allocation by s_i/c_i, subject to the audit floor. Removing the prevention term gives variance-only allocation, proportional to $\frac{d_i}{\sqrt{c_i}}$ when no coordinate is clipped. These endpoints make the joint policy directly comparable with prevention-oriented and precision-oriented designs. The surrogate center is held fixed during this optimization. Choosing $g_i=h_i/2$ minimizes its worst-case error bound, but may increase the actual design variance. Keeping s_i distinct prevents that centering choice from silently changing the prevention forecast. An audit budget also imposes an analytical lower bound on the achievable worst-case variance:

$$\overline{V}_g(q) \geq \frac{\left(\sum_{i=1}^{N} d_i \sqrt{c_i}\right)^2}{B} - \sum_{i=1}^{N} d_i^2 . \qquad (12)$$

This follows from Cauchy–Schwarz applied to $d_i/\sqrt{q_i}$ and $\sqrt{c_i q_i}$ Equality is possible only when the proportional allocation is feasible and exhausts the budget. The bound may be loose when probabilities hit their floor or ceiling. It quantifies a precision cost of limited audit resources without asserting a lower bound on every conceivable confidence interval.

Computational Experiment

The experiment fixes two synthetic populations of 6,000 actions in four strata, detailed in Table 2. All actions have unit audit cost. The first population's defect frequencies agree with its frozen probability scores. The second changes only the final stratum's defect frequency from 0.01 to 0.35 while leaving its score at 0.01. A defective action has loss h_i; a correct action has zero loss. Counts are fixed integers, not newly sampled in each replication. The word "shift" describes this deliberately imposed score–frequency mismatch rather than an observed deployment event.

Table 2: Synthetic Populations and Information Available Before Auditing.

Stratum	Actions	Loss bound	Frozen defect score	Aligned defects	Shifted defects
Routine	3,000	1	0.03	90	90
Ambiguous	1,800	2	0.20	360	360
Consequen-tial	900	5	0.12	108	108
Concealed	300	10	0.01	3	105

The surrogate equals the frozen defect score multiplied by the loss bound. The policies do not receive the defect counts. The expected budget is B=1,200, and the audit floor is q_{min}=0.01. Uniform auditing uses probability 0.20. Risk priority minimizes predicted residual loss with the same floor. The joint allocation uses λ=0.01 in the stated synthetic loss units. Variance-only allocation minimizes $\overline{V}_g$. A stronger score-dependent baseline, termed model variance, minimizes the predicted error variance using $h_i^2\, p_i(1-p_i)$ in place of d_i^2, where p_i is the frozen defect score. Its probabilities are proportional to $h_i\sqrt{p_i(1-p_i)}$, subject to clipping. Table 3 records every resulting probability, so no policy is an unspecified heuristic.

Table 3: Audit Probabilities at the Common Expected Budget.

Policy	Routine	Ambigu-ous	Conse-quential	Concealed
Uniform	0.200000	0.200000	0.200000	0.200000
Risk priority	0.010000	0.148333	1.000000	0.010000
Joint allocation	0.080873	0.154780	0.472113	0.846251
Variance only	0.091509	0.150943	0.415094	0.933962
Model variance	0.055138	0.258580	0.525178	0.321605

Expected residual loss and estimation root mean squared error (RMSE) are calculated exactly using Equations (2) and (6). For computational verification, 20,000 independent audit replications are generated for each of the ten policy–population combinations. Within each stratum, independently sampled binomial counts of audited defective and correct actions are exactly equivalent to individual independent Bernoulli audits because the relevant quantities are constant within each group. NumPy's default random-number generator uses seed 20261016 plus 100 times the population index and the policy index, both starting at zero in the displayed order. The confidence level is 95%.

The primary metrics are residual loss, realized-target RMSE, the untruncated Bernstein radius, and empirical coverage of Equation (9). A supplementary sensitivity grid varies the audit fraction over 0.10, 0.20, 0.30, and 0.40 and the concealed defect frequency over 0.01, 0.10, 0.20, 0.35, and 0.50. A 200-point logarithmic sweep of λ from 10^{-6} to 10^{2} traces the objective trade-off. These are synthetic mechanism checks, not estimates of population performance. No significance tests are used for exact analytical comparisons.

Results

The Uncertainty Target Changes the Answer

An exact diagnostic considers 100 unit-loss defective actions, audit probability 0.8, and g=0. The variance of the estimator around average policy residual loss is 1, whereas its mean squared error for realized residual loss is 25. Using the former standard deviation in a nominal 95% one-sided normal upper bound gives only 65.4% realized-target coverage, calculated exactly from the binomial distribution. Using the correct standard deviation gives 94.4%, reflecting the remaining normal-approximation error. Neither normal interval is the finite-sample certificate in Proposition 2. The diagnostic isolates target confusion from score misspecification, optimization, and approximate simulation.

Prevention and Precision Compete Under Aligned Scores

Table 4 reports exact policy comparisons, and Figure 1 displays the residual-loss results. The model-variance baseline separates conventional score-based variance optimization from the loss-bound protection used by the joint policy. Under aligned scores, risk priority yields the smallest expected residual loss, 732.0 units. The joint allocation leaves 981.0 units, an increase

of 34.0%, because it diverts checks toward actions with large possible surrogate errors. In exchange, its realized-target RMSE falls from 211.3 to 100.0 units. Uniform auditing has both higher residual loss and higher RMSE than the joint policy in this particular population. However, model-variance allocation outperforms the joint policy on both actual metrics under aligned scores, with residual loss 895.6 and RMSE 87.0. The joint policy therefore pays a measurable price for its smaller worst-case uncertainty envelope even against a conventional precision-oriented baseline.

Table 4: Exact Loss and Precision with Empirical Confidence Coverage.

Population and policy	Residual loss	RMSE	Bound radius	Coverage
Aligned: uniform	1,104.0	125.1	1,150.2	100.0%
Aligned: risk priority	732.0	211.3	4,400.0	100.0%
Aligned: joint allocation	981.0	100.0	711.1	100.0%
Aligned: variance only	1,010.9	103.5	704.7	100.0%
Aligned: model variance	895.6	87.0	923.7	100.0%
Shifted: uniform	1,920.0	235.9	1,150.2	100.0%
Shifted: risk priority	1,741.8	1,017.0	4,400.0	100.0%
Shifted: joint allocation	1,137.8	108.7	711.1	100.0%
Shifted: variance only	1,078.3	106.9	704.7	100.0%
Shifted: model variance	1,587.6	169.3	923.7	100.0%

Note: Residual loss is an audit-design expectation. RMSE is the exact square root of Equation (6). Radius is the additive, untruncated quantity in Equation (9). Coverage is based on 20,000 independent audit replications per row; it is not a claim that the theoretical confidence level equals 100%.

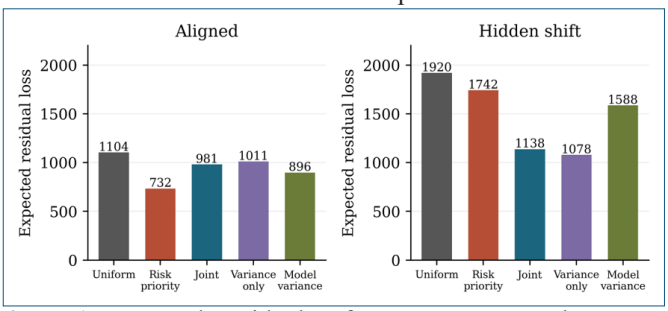


Figure 1: Expected Residual Defect Loss at an Equal Expected Audit Budget. Values are Analytical, So Sampling-Error Bars are not Applicable. The Two Populations Differ Only in the Concealed Stratum's Defect Count.

A Concealed Shift Exposes the Cost of Weak Audit Support

The concealed stratum has the highest loss bound but a low predicted defect frequency. Risk priority audits only 1% of it. After the imposed shift, the policy's residual loss rises to 1,741.8 units and its estimation RMSE reaches 1,017.0 units. The joint allocation audits approximately 84.6% of that stratum and yields 1,137.8 residual units with RMSE 108.7. Relative to risk priority, these correspond to reductions of 34.7% and 89.3%, respectively. The model-variance baseline leaves 1,587.6 loss units with RMSE 169.3 under the shift, compared with 895.6 and 87.0 under aligned scores. Its score-dependent precision criterion does not protect against the concealed mismatch as strongly as the loss-bound criterion. The comparison does not establish universal superiority. Variance-only allocation performs better than the joint allocation on both reported metrics in the shifted population, leaving 1,078.3 loss units with RMSE 106.9. The chosen shift places extra errors in precisely the high-bound stratum emphasized by precision-oriented designs. Under a different loss pattern or misspecified severity bounds, the ranking can change. The benefit is a consequence of the specified population and objective, not an empirical law about AI oversight.

Model assistance also improves precision relative to the unassisted estimator at the same joint allocation: aligned RMSE decreases from 109.4 to 100.0 and shifted RMSE from 117.5 to 108.7. The effect is modest compared with the allocation effect and should not be conflated with it. The uncorrected surrogate estimate of remaining loss is strongly biased after the shift. Its empirical mean error is approximately −1,009.8 units under risk priority and −156.6 units under joint allocation. By comparison, the model-assisted estimator's empirical mean errors are 4.55 and −0.12 units, with Monte Carlo standard errors 7.23 and 0.77. This illustrates why targeted checking and confident surrogate predictions cannot substitute for probability-weighted residual correction.

The Confidence Bounds are Valid but Conservative

All ten experiments show zero coverage failures in 20,000 replications. For each row separately, a one-sided 95% exact binomial lower confidence limit for the simulation coverage probability is approximately 99.985%. This diagnostic supports the implementation but does not improve the theorem's stated 95% guarantee or establish calibration outside the specified populations.

A centering ablation holds the joint allocation fixed and replaces g_i with $h_i/2$, leaving the prevention forecast unchanged. In the aligned population, this lowers the radius from 711.1 to 400.4 units but increases actual RMSE from 100.0 to 161.3. A smaller worst-case envelope is therefore not evidence of better actual precision. The confidence radii are much larger than the actual RMSE values. For risk priority, the 4,400.0-unit radius exceeds even the entire shifted population's pre-audit loss of 2,400.0. The joint radius is smaller at 711.1 units but remains large. These results expose the cost of protecting against every loss configuration permitted by the bounds. Reporting only coverage would obscure this lack of precision. The method provides an honest certificate when useful; it does not guarantee that a limited audit budget can produce a useful certificate.

Numerical Verification and the Allocation Frontier

Enumeration of all 256 audit selections in an eight-action heterogeneous example produces mean estimation error below 10^{-12} in absolute value. The enumerated mean squared error, 14.1096428571, matches Equation (6). An independent

constrained numerical optimizer matches the closed-form allocation objective within 10^{-7} in a heterogeneous-cost example. These checks examine the estimator identity and allocation rule separately from the larger simulation. Figure 2 plots predicted residual loss against the square root of the worst-case variance envelope. Moving toward stronger precision protection sacrifices prevention under the frozen score model. Uniform allocation lies above the achievable trade-off in this example. The curve concerns the proxy objective and its variance envelope, not the unknown true-loss frontier. The sensitivity outputs provide the corresponding exact losses and RMSE values under alternative budgets and concealed-error frequencies. At the shifted error frequency of 0.35, increasing the expected audit fraction from 0.10 to 0.40 reduces joint-policy residual loss from 1,746.5 to 481.8 units and RMSE from 191.9 to 45.0. These gains depend on the synthetic severity structure and do not identify a universally preferred budget.

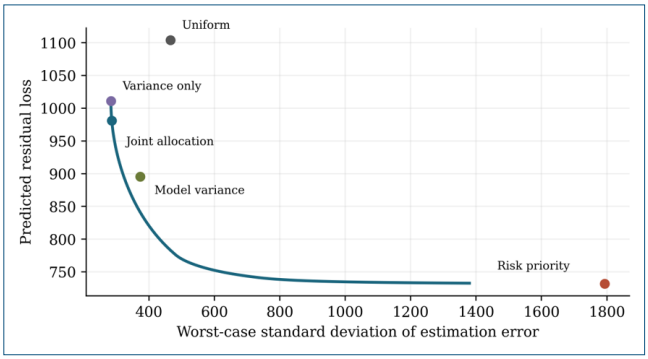


Figure 2: The Prevention and Observability Trade-Off for the Fixed Surrogate Model and Expected Budget of 1,200 Audits. The Curve Varies the Precision Penalty; Labeled Points Show the Five Evaluated Policies. Smaller Values on Either Axis are Preferable, Subject to the Trade-Off.

Discussion

Audits are Part of the Inferential Design

An audit log records the cases an operator chose to inspect. It becomes evidence about the uninspected cases only through additional assumptions or a sampling design. The present framework makes that dependency explicit. A system can correct many errors while producing weak evidence about the errors it did not inspect. A low observed defect count is particularly uninformative when large parts of the action space have little chance of being checked. The practical implication is to store the propensity, loss bound, frozen surrogate, audit outcome, and correction status for each proposal. Aggregate audit volume alone is insufficient. A dashboard should distinguish predicted loss, estimated residual loss, and its confidence bound. It should also distinguish a statistically valid but large bound from evidence that a release tolerance has been met. These distinctions give operational meaning to runtime assurance and verification debt without relying on their broader architectural claims [11,13].

What the Method Adds and What Remains Classical

The contribution is the integration of corrective sampling and the random residual target. The estimator remains a generalized difference estimator followed by subtraction of observed corrections. The proof uses classical inverse-probability identities; the confidence bound uses Bernstein concentration; and the allocation uses convex optimality conditions. These foundations strengthen rather than diminish the result: they expose which claims depend on new problem formulation and which rely on established mathematics. The frontier formalizes an operational choice that single-purpose evaluation does not settle. A probability design preferred for preventing expected defects need not be preferred for learning the remaining total. Conversely, worst-case precision protection may spend heavily on severe but rare possibilities. The correct allocation depends on the value assigned to prevention, evidence, and the credibility of the specified loss bounds. The model gives an explicit decision surface, not a universal prescription for an optimal audit fraction.

Implementation Boundaries

Independent Bernoulli auditing controls cost in expectation. It does not enforce a hard cap on the realized number of audits. An operator who stops auditing when capacity is exhausted changes the sampling design and cannot retain the original probabilities or proofs. A deployable implementation must reserve capacity, use an appropriate fixed-size design with its own variance analysis, or redesign the protocol. The manuscript's comparisons hold at an equal expected budget, not at an exactly equal realized audit count. Similarly, the estimator is not justified if the surrogate is fitted to the same audit labels and then substituted retrospectively. Historical training, an independent pilot, or an appropriately proved sequential or sample-splitting scheme is required. Across repeated batches, a new frozen design can be formed from earlier evidence, but the present confidence statement is per batch. Monitoring many batches or choosing among many post-audit policies requires simultaneous or sequential error control.

Perfect verification and complete correction are substantive assumptions. If a checker misses defects, the observed label is not y_i. If a correction leaves residual damage or introduces a new defect that is not measured, the subtracted corrected total is wrong. There is a simple extension when the post-correction losses of audited actions are measured exactly: add their observed total to both L and $\hat{L}_g$, and add the same total to U_g. The estimation error is unchanged, so Propositions 1 and 2 continue to hold for total post-audit loss. This extension does not solve inaccurate auditing or unknown remediation outcomes. Modeling verifier sensitivity, dependence, and remediation efficacy is therefore a separate extension. Hu et al. [10] provide a relevant precedent for explicitly handling imperfect verification, but their guarantee cannot simply be imported into the present estimator.

Limitations and a Falsifiable Validation Path

The study contains no benchmark of a trained classifier or language-model agent. Its synthetic strata intentionally isolate heterogeneous severity and hidden score error. They establish that the stated mechanism can occur, not how often it occurs. The chosen penalty is illustrative, and the conservative variance envelope does not use information about aggregate defect prevalence. Sharper bounds, cost-sensitive labels, cluster audits, or externally validated restrictions on the loss vector may improve practical performance. A substantive follow-up should freeze a collection of AI-generated actions for which independent labels are available, such as schema-constrained data transformations or executable program changes. Surrogates should be trained on disjoint material. All proposals should receive offline reference adjudication for evaluation, while each simulated audit policy sees only its sampled labels. The evaluation should compare exact-budget and expected-budget designs, report prevention loss, confidence-bound width, failure frequency, verifier error, and latency, and reserve entire task families for shift testing.

The present policy would be challenged if its precision gains disappeared after accurate costs and loss bounds were introduced, or if its extra prevented loss under shift failed to compensate for its aligned-score cost.

Severity weights also encode values that statistical validity cannot justify. A precisely estimated weighted loss can conceal a concentrated burden on a subgroup. Group-specific bounds and minimum audit allocations can be added, but their choice requires domain judgment, and simultaneous confidence statements require an explicit multiplicity adjustment. Statistical observability should support accountable decisions, not define their acceptable consequences. Audit labels should be collected before any harmful release, and the statistical design does not justify experimenting with unbounded or prohibited actions.

Conclusion

Corrective AI audits change the population of defects that remains to be assessed. For this setting, the uncertainty of an estimator around average policy risk can be far smaller than its uncertainty about the actually unaudited batch. A model-assisted residual estimator, a finite-sample certificate, and a convex allocation rule make this difference operational. The synthetic results show both the benefits and the price of reserving audit capacity for observability. They also show that formally valid confidence bounds can be too conservative to certify a demanding release threshold. Effective assurance therefore requires an explicit statistical target, a defensible audit design, and a transparent decision about how much prevention to exchange for evidence.

Funding Statement: No external funding was received for this study.

Conflict of Interest Statement: The author declares no competing interests. Three cited publications are earlier works by the author.

Data and Code Availability: All populations are synthetic and specified in Table 2. Supplementary Material S1 accompanies the manuscript and contains the executable reproduction script, numerical outputs, sensitivity results, and original figure files. No private or third-party dataset is required. The script regenerates all reported numerical comparisons and computational checks.

Ethical Approval: The study involved no human participants, animal subjects, personal data, or real-world interventions. Ethical approval was not applicable.

References

1. Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 275-284. DOI: https://doi.org/10.1145/3097983.3098066.
2. Kossen, J., Farquhar, S., Gal, Y., Rainforth, T. (2021). Active testing: Sample-efficient model evaluation. In International Conference on Machine Learning, PMLR, 139, 5753-5763. https://proceedings.mlr.press/v139/kossen21a.html.
3. Kossen, J., Farquhar, S., Gal, Y., Rainforth, T. (2022). Active surrogate estimators: An active learning approach to label-efficient model evaluation. Advances in Neural Information Processing Systems, 35, 24557-24570. DOI: https://doi.org/10.52202/068431-1783.
4. Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685. DOI: https://doi.org/10.1080/01621459.1952.10483446.
5. Särndal, C.E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. Springer, https://link.springer.com/book/9780387406206.
6. Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T. (2023). Prediction-powered inference. Science, 382(6671), 669-674. DOI: https://doi.org/10.1126/science.adi6000.
7. Ashouritaklimi, K., Kilian, V., Huang, D., Rainforth, T., Caron, F. (2026). Prediction-Powered Active Testing. arXiv preprint arXiv:2607.08347, https://arxiv.org/pdf/2607.08347.
8. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T. (2024). Conformal risk control. In International conference on learning representations, https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.
9. Manoharan, J. (2026). Audited Selective Verification for Risk-Controlled N-1 Thermal Contingency Screening under Deployment Shift. arXiv preprint, https://arxiv.org/abs/2607.13221.
10. Hu, M., Hu, S., Guo, X., Wang, X., Wang, B., et al. (2026). Audit-First VAPO: Risk-Certified Selective Updates under Imperfect Verification. arXiv preprint, https://arxiv.org/abs/2609.33662.
11. Tan, K.H. (2026). Runtime assurance for enterprise agentic AI systems: A policy-gated control model with quantitative autonomy-risk scoring. World Journal of Advanced Research and Reviews, 31(1), 512-522. DOI: https://doi.org/10.30574/wjarr.2026.31.1.1872.
12. Tan, K.H. (2026). Containing cascading AI errors: A spectral branching-process framework for verification allocation in multi-agent systems. International Journal of Progressive Research in Engineering Management and Science, 6(8), 374-383. DOI: https://doi.org/10.58257/IJPREMS54572.
13. Tan, K.H. (2026). The coordination compression trap: A dynamic model of AI-enabled productivity, verification debt, and organisational resilience in global knowledge firms. Journal of Global Economics, Management and Business Research, 18(3), 256-278. DOI: https://doi.org/10.56557/jgembr/2026/v18i310956.
14. Boucheron, S., Lugosi, G., Massart, P. (2013). Concentration inequalities: A nonasymptotic theory of independence. Oxford University Press, https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.
15. Boyd, S., Vandenberghe, L. (2004). Convex optimization. Cambridge University Press, https://web.stanford.edu/~boyd/cvxbook/.

Copyright: ©2026 Kwan Hong Tan. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.