THE MODULARITY TRAP IN PRIVATE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Orthogonal scores remove the first-order effect of nuisance estimation; differential privacy adds a second operation, bounding the contribution of each record. We show that where the bound is placed matters: clipping the completed orthogonal score can change the nuisance rates required for inference, even when the clip keeps the null mean exactly zero. No bounded nonlinear scalar map escapes the effect, since preserving every centered nuisance tangent forces affinity. An exact residual-product construction makes the consequence sharp. With nuisance errors of order $m^{-\gamma}$ for $1/4 < \gamma < 1/2$, the inherited zero-centered test on the clipped product rejects the null with probability tending to one, while the Prediction-Bounded Orthogonal Score (PBOS), bounding a centered residual multiplier instead of the product, keeps size α under the same rates and admits finite-sample calibration with a lattice release. The rate gap resurfaces as a threshold gap once each method is recalibrated, and the threshold gap as a power gap: in text-representation audits under strong confounding, both placements retain nearly identical standardized signal, yet PBOS attains 92.27% power against 73.93% for product clipping.

1 INTRODUCTION

Orthogonal estimation controls how errors in fitted nuisance functions enter a target equation (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023). At the null, the derivative of its mean along a nuisance path vanishes, so the remaining error can be a product of estimation errors. Differential privacy introduces a second operation: bound the contribution of each record before releasing a noisy sum. For a residual-product score, clipping the completed product supplies such a bound. Whether its nuisance cancellation survives this operation determines which estimation rates suffice for inference.

The issue arises even when clipping preserves the null mean. Differentiating the transformed score weights each nuisance contribution by the derivative of the clipping map. Contributions in the saturated region disappear; those in the interior remain. Their original cancellation can be lost. Exact centering at one fitted value consequently leaves the response to fitting error unresolved.

We characterize this obstruction for every locally Lipschitz scalar transformation: preserving all centered nuisance tangents forces affinity (Theorem 1). A finite-support residual model then makes the inferential consequence exact. Clipping the product turns the product remainder of orthogonalization into a first-order one; placing the bound on the centered multiplier keeps it second order. With nuisance errors of order $m^{-\gamma}$ for $1/4 < \gamma < 1/2$, the inherited zero-centered test on the clipped product rejects the null with probability tending to one, while PBOS keeps its nominal level (Theorem 3). Recalibrating against the fitted null restores size for both tests, but the first-order shift moves into the critical value: Figure 1 shows product clipping paying a growing threshold where PBOS remains at the normal quantile.

The construction rests on the binary label residual: multiplied by a bounded function of the representation residual, it is centered by conditional expectation before any differentiation is needed, and centering the bounded multiplier removes the prediction tangent as well. This is the Prediction-Bounded Orthogonal Score (PBOS). The same factorization shortens the conditional label span, which shrinks the concentration term in a finite-sample rejection threshold; calibration absorbs the realized fitting shift.

Representation audits give the practical consequence. A query direction can reflect both a target concept and nuisance context (Kim et al., 2018; Hewitt et al., 2021; Kumar et al., 2022). On a fixed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

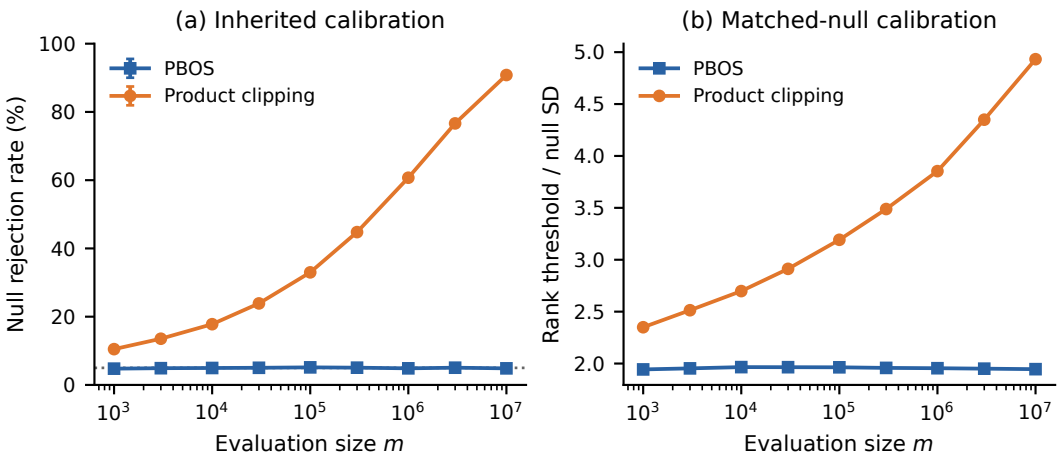


Figure 1: Score placement changes testing validity and calibration cost. In the controlled null model of Section 4.2, both scores have the same nuisance errors. Left: inherited zero-centered tests under $t_m = s_m = 0.2m^{-1/3}$, using 50,000 paired runs per point and Wilson 95% intervals. The dotted line is 5%. Right: matched-null absolute-statistic rank thresholds divided by the true null standard deviation, using 20,000 calibration draws per method and sample size. Independent validation of these thresholds returns both tests to approximately 5% size (Appendix F).

residual query, the two placements retain similar standardized signal. Their calibrated thresholds differ, explaining the power and sample-size gains measured below.

2 Related work

The residual-product starting point comes from orthogonal estimation and conditional testing. Double machine learning and locally robust estimation control the first-order effect of nuisance fitting (Chernozhukov et al., 2018; 2022); orthogonal statistical learning develops the corresponding learning guarantees (Foster & Syrgkanis, 2023). The generalized covariance measure uses a product of residuals for conditional testing, and its weighted extension enlarges the alternatives detected by that moment (Shah & Peters, 2020; Scheidegger et al., 2022). Our question concerns the transformation of such a score after its cancellation has been established. Concept directions and conditional probes supply the representation queries used in our experiments (Kim et al., 2018; Hewitt et al., 2021).

Bounding an estimating equation connects robust statistics to differential privacy (Dwork & Lei, 2009; Avella-Medina, 2021). Clipping bias and the trade-offs among bias, accuracy, and privacy have been studied directly (Avella-Medina, 2022; Kamath et al., 2025). Our construction keeps the null mean exactly zero and changes its nuisance derivative. This distinguishes the source of the shift that must be calibrated. Analyses of private tests likewise depend on the statistic: examples include simple-hypothesis tests, regression tests, and conditional-independence tests (Canonne et al., 2019; Alabi & Vadhan, 2023; Kalemaj et al., 2024). Conditional-independence testing is the closest setting; we isolate how bounding placement changes the residual moment and its calibration cost.

DP-CATE privatizes orthogonal treatment-effect learners by output perturbation (Schröder et al., 2025), and private causal workload design uses orthogonal moments to preserve inferential targets (Asiaee & Aryan, 2026). These works motivate designing the release around the estimating equation. Here the operation under study is a bounded scalar transformation of the score itself. Private calibration and bootstrap methods analyze the distribution of a privatized procedure (Kazan et al., 2023; Wang et al., 2025); our calibration measures the null shift generated by the full nuisance-fitting procedure.

3 PROBLEM FORMULATION

Each record consists of a binary target $Y \in \{0, 1\}$, a real-valued score Z along a prespecified representation direction, and observed nuisance context N. The direction and feature construction are fixed before evaluation. Under a population law P, write $q_P(N) = \mathbb{E}_P[Y \mid N]$ and let $g_P(N)$ be a residual center. Training and evaluation use disjoint records. The fitted probability $p(N) \in [0, 1]$ and residual center $\widehat{g}(N)$ are frozen before evaluation labels are drawn.

For an odd, nondecreasing, Lipschitz function d bounded in absolute value by $B > 0$, define

$$\mu_d(P) = \mathbb{E}_P[\{Y - q_P(N)\}d\{Z - g_P(N)\}]. \tag{1}$$

We seek evidence of a nonzero moment along the chosen residual direction. Conditional independence $Y \perp Z \mid N$ implies $\mu_d(P) = 0$, and the moment targets directional residual association (Appendix A). For local analysis and calibration, fix a reference null law P_0 and write $q_0 = q_{P_0}$, $g_0 = g_{P_0}$, and $A_0 = Z - g_0(N)$; the signal paths below hold these anchors fixed while changing the label law.

Under a fixed law P, let ψ_t denote a score evaluated along a nuisance path, with derivative $\dot{\psi} = \partial_t \psi_t|_{t=0}$. Centering requires $\mathbb{E}\psi_0 = 0$. Orthogonality along the path requires $\partial_t \mathbb{E}\psi_t|_{t=0} = 0$. All derivative statements below use domination sufficient to interchange differentiation and expectation. We write $\|f\|_2 = (\mathbb{E}f^2)^{1/2}$ and $\mathrm{clip}_B(u) = \max\{-B, \min(u, B)\}$.

Two datasets are adjacent when one record is replaced in a fixed training or evaluation slot. A release M is (ε, δ)-differentially private if $P\{M(D) \in A\} \le e^{\varepsilon} P\{M(D') \in A\} + \delta$ for every adjacent pair and measurable event A (Dwork & Roth, 2014). Our objective is a private noisy sum and a two-sided rejection rule with null rejection probability at most α. The construction must control both the response to record replacement and the null shift caused by nuisance fitting.

4 SCORE PLACEMENT AND PRIVATE INFERENCE

The argument starts with the derivative of a transformed score. Its expectation determines the nuisance remainder, and the order of that remainder determines whether a centered critical value remains valid. The PBOS construction follows by preserving the cancellation used in this argument.

4.1 THE MODULARITY TRAP

Consider a Lipschitz scalar transformation b. Assume it is differentiable at ψ_0 almost surely and that $|\psi_t - \psi_0|/|t|$ has an integrable envelope near zero. The chain rule and dominated convergence give

$$\partial_t \mathbb{E}b(\psi_t)|_{t=0} = \mathbb{E}[b'(\psi_0)\dot{\psi}]. \tag{2}$$

The derivative of the transformation reweights the nuisance tangent. Even if its original mean vanishes, the reweighted mean can be nonzero. For clipping, away from probability mass at the knots $\{-B, B\}$, the weight is $\mathbf{1}\{|\psi_0| < B\}$.

Requiring this cancellation for every centered path imposes a sharp restriction on the transformation. Local Lipschitz continuity is enough; it includes both hard clipping and smooth bounds.

Theorem 1 (Universal preservation of centered nuisance tangents)**.** *Let $b : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz. Suppose that for every finitely supported random pair (S, H) with $\mathbb{E}S = \mathbb{E}H = 0$, at whose score support values b is differentiable,*

$$\partial_t \mathbb{E}b(S + tH)|_{t=0} = 0. \tag{3}$$

Then b is affine. If b is globally bounded, it is constant.

Proof sketch. For any two differentiability points x, y, choose a centered distribution supported on x, y and, if needed, a third point, assigning opposite tangents to x, y and zero tangent to the third point. Equation (3) forces $b'(x) = b'(y)$. Thus b' is constant almost everywhere, and local absolute continuity gives affinity. The full construction is in Appendix B. □

The quantifier over score paths does the work here: it lets the proof compare the derivative of b at any two differentiability points. For a specified score, the same calculation gives an exact criterion:

$$\partial_t \mathbb{E} b(\psi_t)|_{t=0} = \mathrm{Cov}\{b'(\psi_0), \dot{\psi}\} \quad \text{when } \mathbb{E}\dot{\psi} = 0. \tag{4}$$

A constant derivative on the score support makes this covariance zero. The residual-product construction below shows how a nonzero covariance changes the rate required for inference.

4.2 From nuisance order to testing validity

Let $P(Z = 1) = 2/3$, $P(Z = -2) = 1/3$, and let $Y \sim \mathrm{Bernoulli}(1/2)$ independently of Z. Then $\mathbb{E}Z = 0$. For prediction error t and residual-center error s, compare

$$\begin{aligned} \psi^{t,s}_{\mathrm{raw}} &= (Y - 1/2 - t)(Z - s), \\ \psi^{t,s}_{\mathrm{post}} &= \mathrm{clip}_{3/4}\{(Y - 1/2 - t)(Z - s)\}, \\ \psi^{t,s}_{\mathrm{PB}} &= (Y - 1/2 - t)\,\mathrm{clip}_{3/4}(Z - 5/8 - s). \end{aligned} \tag{5}$$

The PBOS center $5/8$ solves $\mathbb{E}\,\mathrm{clip}_{3/4}(Z - g) = 0$. Thus the two residual centers are defined by their respective population centering equations, with the same perturbation s.

Proposition 2 (An exact nuisance-order gap)**.** *For $|t| \le 1/16$ and $|s| \le 1/8$, the scores in Equation* (5) *satisfy*

$$\mathbb{E}\psi^{t,s}_{\mathrm{raw}} = ts, \qquad \mathbb{E}\psi^{t,s}_{\mathrm{post}} = -\frac{2}{3}t(1 - s), \qquad \mathbb{E}\psi^{t,s}_{\mathrm{PB}} = \frac{2}{3}ts. \tag{6}$$

Both $Z = -2$ label branches saturate after product clipping. They contribute no local prediction derivative, leaving the term linear in t in Equation (6). For PBOS, the bounded multiplier has mean $-2s/3$, so multiplication by the mean label residual $-t$ gives $2ts/3$. Thus all three scores have the same zero mean at the anchor, while their responses to estimation error differ in order. Appendix D computes the branches and variances.

Theorem 3 (Same nuisance rates, different testing validity)**.** *Draw m iid observations from this model and take deterministic $t_m, s_m \to 0$. For $j \in \{\mathrm{post}, \mathrm{PB}\}$, let $\widehat{v}_{m,j}$ consistently estimate the score variance and let independent release noise satisfy $W_m = o_P(\sqrt{m})$. Define*

$$U_{m,j} = \frac{\sum_{i=1}^{m} \psi^{t_m, s_m}_{j,i} + W_m}{\sqrt{m\widehat{v}_{m,j}}}.$$

If $\sqrt{m}|t_m| \to \infty$ and $\sqrt{m}|t_m s_m| \to 0$, then for $0 < \alpha < 1$ and standard normal quantile $z_{1-\alpha/2}$,

$$P\{|U_{m,\mathrm{post}}| > z_{1-\alpha/2}\} \to 1, \qquad P\{|U_{m,\mathrm{PB}}| > z_{1-\alpha/2}\} \to \alpha. \tag{7}$$

Proof sketch. The centered sums obey a bounded-array central limit theorem, with limiting variances $17/48$ and $9/128$. What separates the tests is their standardized mean: Equation (6) makes it diverge for product clipping and vanish for PBOS. Consistent variance estimation and $W_m = o_P(\sqrt{m})$ preserve these limits by Slutsky's theorem. Appendix D gives the full argument. □

For $t_m = c_q m^{-\gamma_q}$ and $s_m = c_g m^{-\gamma_g}$ with positive exponents and nonzero constants, the strict separation region is

$$\gamma_q < \frac{1}{2}, \qquad \gamma_q + \gamma_g > \frac{1}{2}. \tag{8}$$

Equal rates give $1/4 < \gamma < 1/2$, connecting score placement to the familiar product-rate condition in orthogonal inference (Chernozhukov et al., 2018). This result concerns inherited zero-centered calibration. Matched-null recalibration absorbs the induced first-order shift into its threshold, as the experiments below demonstrate.

4.3 PREDICTION-BOUNDED ORTHOGONAL SCORES

With $\Lambda(u) = (1 + e^{-u})^{-1}$, fitted probability p, and representation residual $a = Z - \widehat{g}(N)$, define

$$\phi_d(Y; p, a) = (Y - p)d(a), \qquad |d(a)| \le B. \tag{9}$$

We use $d = \mathrm{clip}_B$ in the experiments. The binary residual form follows from a simple centering identity: every score satisfying $p\, s(1; p, a) + (1-p)s(0; p, a) = 0$, with $0 < p < 1$, can be written as $(Y - p)c(p, a)$. PBOS selects a bounded multiplier independent of p. This keeps the label centering outside the nonlinear transformation.

Let P_θ keep the marginal law of (Z, N) fixed and change the conditional label mean $m_\theta(Z, N) = \mathbb{E}_\theta[Y \mid Z, N]$, with $m_0 = q_0(N)$. For nuisance paths $q_t = q_0 + th_q$ and $g_s = g_0 + sh_g$, write

$$\mathcal{M}(\theta, t, s) = \mathbb{E}_\theta[\{Y - q_t(N)\}d\{Z - g_s(N)\}].$$

Theorem 4 (PBOS nuisance and signal derivatives)**.** *Assume h_q is integrable, the paths q_t are valid probabilities, and differentiation of m_θ is dominated under expectation. At the conditional null,*

$$\mathcal{M}(0, 0, s) = 0 \quad \textit{for every admissible } s, \tag{10}$$

$$\partial_t \mathcal{M}(0, 0, 0) = -\mathbb{E}[h_q(N)d(A_0)], \tag{11}$$

$$\partial_\theta \mathcal{M}(0, 0, 0) = \mathbb{E}[\dot{m}_0(Z, N)d(A_0)], \tag{12}$$

where $\dot{m}_0 = \partial_\theta m_\theta|_{\theta=0}$.

Proof sketch. Condition on (Z, N) and replace Y by $m_\theta(Z, N)$. At the null, the label residual has mean zero for every g_s. Differentiating the remaining two paths gives the identities; Appendix C contains the complete proofs and the logistic specialization. □

Residual-fitting tangents vanish along the whole null path, including at clipping knots. Prediction tangents vanish for every integrable h_q when

$$\mathbb{E}[d\{Z - g_0(N)\} \mid N] = 0. \tag{13}$$

For clipping, this is a conditional Huber location equation. It centers the bounded residual, whereas a conditional mean only centers the untransformed residual. A parametric prediction model needs the corresponding moment condition only along its feature tangent space. In particular, logistic prediction with features $v(N)$ requires $\mathbb{E}[q_0(1 - q_0)v(N)d(A_0)] = 0$.

The same condition controls finite fitting errors. If d is L_d-Lipschitz and the errors are square-integrable, conditional on independent training,

$$\left|\mathbb{E}[\{Y - q(N)\}d\{Z - g(N)\}]\right| \le L_d \|q - q_0\|_2 \|g - g_0\|_2. \tag{14}$$

This follows by subtracting $d(A_0)$ and applying Cauchy–Schwarz. If Equation (13) has residual $r(N)$, add $\|q - q_0\|_2\|r\|_2$. For the implemented regularized fits, calibration includes the complete realized fitting shift and centering error.

On the logistic alternative $m_\theta = \Lambda\{\mathrm{logit}\, q_0(N) + \theta Z\}$, where $\mathrm{logit}(p) = \log\{p/(1-p)\}$, let $\rho_0 = q_0(1 - q_0)$ and assume $0 < q_0 < 1$. If $\mathbb{E}[\rho_0 g_0(N)d(A_0)] = 0$, Equation (12) reduces to $\mathbb{E}[\rho_0 A_0 d(A_0)] \ge 0$. It is positive whenever $A_0 d(A_0) > 0$ with positive probability. The bounded residual therefore retains a directional signal whose strength depends on the variation left after removing context.

For fixed (p, a), the two PBOS values are $(1 - p)d(a)$ and $-p\, d(a)$. Their distance is $|d(a)| \le B$. Product clipping, $\mathrm{clip}_B\{(Y - p)a\}$, can instead attain both B and $-B$, giving span $2B$. Hoeffding's lemma assigns variance proxy one quarter of the squared span (Hoeffding, 1963). PBOS thus reduces the worst-case label contribution from B^2 to $B^2/4$ per record. Record-replacement sensitivity stays at $2B$ for both scores under a frozen fit.

4.4 A CALIBRATED PRIVATE TEST

Fit the nuisance functions on training records, evaluate PBOS on m disjoint records, and add independent privacy noise. The implemented statistic has the form

$$T^{\mathrm{impl}} = \sum_{i=1}^{m} (Y_i - p_i)d(a_i) + W + E_{\mathrm{sc}}, \tag{15}$$

where $p_i = p(N_i)$, $a_i = Z_i - \widehat{g}(N_i)$, W is the release noise, and E_{sc} is numerical error. Let $\mathcal{F}_{\mathrm{pre}}$ contain the fitted functions, evaluation covariates, public calibration output, and mechanism settings fixed before evaluation labels are generated.

We calibrate the fitted conditional null mean by repeating the full training and fitting procedure under the specified null law. If run k has null probabilities q_{ki}, define

$$R_k = \left| \sum_{i=1}^{m} (q_{ki} - p_{ki}) d(a_{ki}) \right|, \qquad \widehat{B} = \max_{1 \le k \le K} R_k. \tag{16}$$

The synthetic experiments make these probabilities available. The future reference shift is $R_{K+1} = |\sum_i (q_i - p_i) d(a_i)|$. Under exchangeability with the calibration shifts, it exceeds $\widehat{B}$ with probability at most $1/(K+1)$. This rank argument covers fitting randomness without an asymptotic approximation.

Theorem 5 (Finite-sample calibration)**.** *Assume the shifts $R_1, \ldots, R_{K+1}$ are exchangeable. Given $\mathcal{F}_{\mathrm{pre}}$, suppose evaluation labels are independent Bernoulli variables with means $q_i^\star$, the noise W is centered and σ-sub-Gaussian and is independent of the labels, and $|E_{\mathrm{sc}}| \le e$. Let q_i denote the calibration-null probabilities and assume $|\sum_i (q_i^\star - q_i) d(a_i)| \le D_0$. Allow total-variation discrepancy at most τ between the analyzed joint law and its implementation. For $\alpha_c = \alpha - 1/(K+1) - \tau > 0$, set*

$$c_\alpha = \widehat{B} + D_0 + e + \sqrt{2 \left(\frac{mB^2}{4} + \sigma^2 \right) \log \frac{2}{\alpha_c}}. \tag{17}$$

Then $P\{|T^{\mathrm{impl}}| > c_\alpha\} \le \alpha$ jointly over calibration and a future null run.

Proof sketch. Decompose the statistic into the fitted null shift, the allowed probability mismatch, the centered label sum, noise, and numerical error. The conditional label sum has variance proxy at most $mB^2/4$; independent noise adds σ^2. A two-sided sub-Gaussian bound, the calibration rank bound, and the discrepancy allowance give the result. Appendix E carries out the bound. □

With $K = 5{,}000$, the observed maximum shift is $\widehat{B} = 0.094285$. Using $m = 16{,}000$, $B = 0.08$, $\sigma = 1$, $D_0 = 0.30$, $e = 0.01$, $\alpha = 0.05$, and $\tau = 0$ for the target sampling law gives concentration term 14.0165 and threshold 14.4208. We round upward to 14.43 before validation. The guarantee averages over calibration and a future run from that null law.

PBOS also supports finite-sample testing from high-probability error bounds. Suppose, with training probability at least $1 - \beta$, the prediction and residual errors have L^2 bounds a_q, a_g, and the bounded-root defect has bound a_r. Set $\Delta = L_d a_q a_g + a_q a_r$ and $x = \log\{2/(\alpha - \beta)\}$, with $\beta < \alpha$. Appendix D proves that the threshold

$$c_{\mathrm{env}} = m\Delta + e + \sqrt{2\{mB^2(1/4 + a_q^2) + \sigma^2\}x} + \frac{2B}{3}x \tag{18}$$

controls size for iid evaluation records independent of training. This route uses nuisance-error envelopes in place of recordwise null probabilities. The appendix also bounds the prediction-tangent defect through regularization, optimization, generalization, and fitting perturbations.

The nominal representation experiments use nuisance budgets $(3, 1)$ and disjoint evaluation budget 4, giving $(4, 10^{-6})$ privacy in the continuous Gaussian model under its component assumptions. Appendix I establishes a lattice release of the complete fitted sum using exact discrete-Gaussian noise (Canonne et al., 2020). Its solver and arithmetic conditions support the same target at physical noise scale 5.12, with a corresponding statistical threshold. The nominal representation curves use scale 1.

5 Experiments

The experiments follow the two implications of the theory. The finite-support model tests the nuisance-rate prediction directly. Representation audits then measure the power difference when each fitted score receives its own calibration.

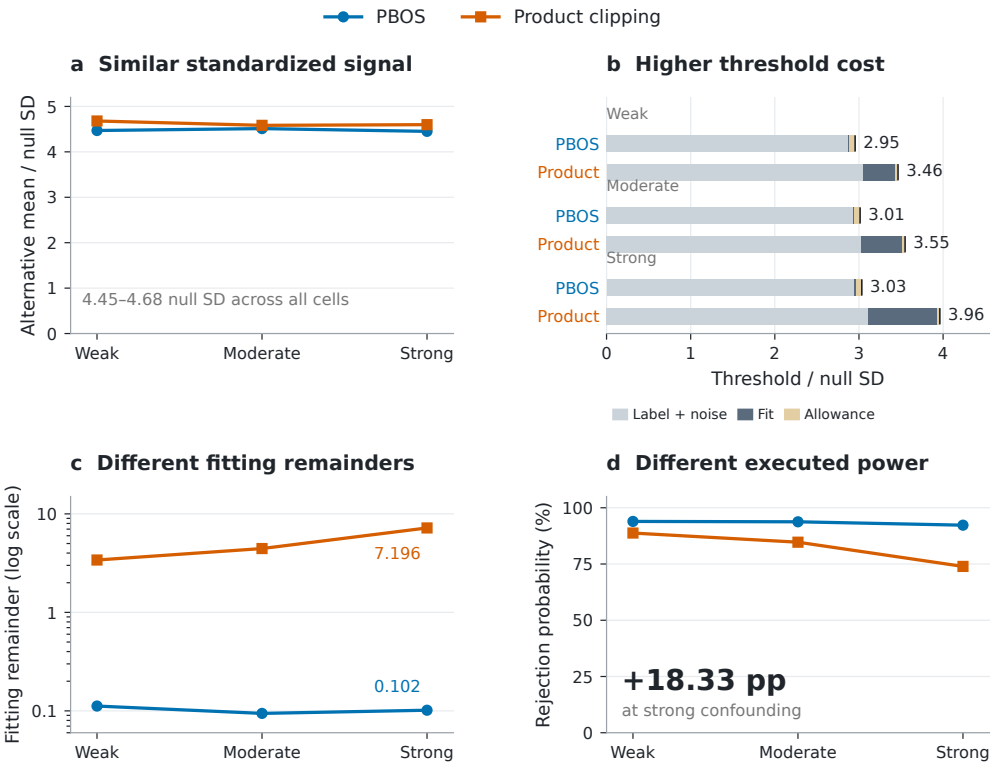


Figure 2: Score placement changes the cost of a calibrated test. At $\theta = 0.20$, (a) alternative means are divided by method-specific null standard deviations; (b) thresholds use the same standardization, with ticks for rounded execution thresholds; (c) fitting shifts are in original score units; and (d) power uses 3,000 runs per cell and Wilson 95% intervals. Null standard deviations use 5,000 runs. The fixed allowance is $D_0 + e = 0.31$.

5.1 Validity separation and recalibration cost

For the model in Section 4.2, take $t_m = s_m = 0.2m^{-1/3}$, $W_m \sim N(0, 4)$, and $\alpha = 0.05$. Nine sample sizes range from 10^3 to 10^7, with 50,000 paired replications per point. Multinomial sampling of the four support states generates the exact iid sum distribution. Both methods use the same nuisance perturbations, records, and release noise; studentization includes sample variance and the known noise variance.

Figure 1 shows the predicted separation. Product clipping's null rejection rate rises from 10.49% at $m = 10^3$ to 90.81% at $m = 10^7$; PBOS remains near 5%, with 4.86% at the largest sample size. The common-rate and asymmetric-rate grids in Appendix F follow the boundaries in Equation (8).

Matched-null recalibration translates the first-order shift into a threshold cost. We select an absolute-statistic order statistic from 20,000 null draws and validate on 30,000 fresh draws per method and size. Both tests recover approximately 5% rejection. Across the sample-size range, the product threshold grows from 2.350 to 4.932 null standard deviations; PBOS stays near 1.96. The same shift that breaks inherited calibration becomes a critical-value tax once recalibrated.

5.2 Score placement on text representations

We use public Wikipedia, Stack Exchange, and PMC Open Access text encoded by Qwen3-Embedding-0.6B (Zhang et al., 2025), with EmbeddingGemma-300M as a second representation (Vera et al., 2025). A frozen benefit-harm query defines Z, and nuisance features capture surface

variation, certainty, and subjectivity; synthetic logistic labels then supply a directional effect θ together with the known calibration-null probabilities. Each primary run uses $16{,}000$ records per split, $B = 0.08$, and the nominal privacy schedule above, with the protocol in Appendix G.

The main comparison is between PBOS, $(Y - p)\,\mathrm{clip}_B(a)$, and product clipping, $\mathrm{clip}_B\{(Y - p)a\}$. Each score receives its own fitting-shift calibration and rejection threshold. We report null rejection rates and power, with two-sided Wilson 95% intervals (Wilson, 1927). Signal summaries divide alternative means by each method's empirical null standard deviation; this descriptive scaling leaves the executed tests unchanged. An unresidualized multiplier is included to examine which directional information residualization removes.

The confounding ladder fixes $\theta = 0.20$ and varies the nuisance contribution to the label model. Each method and setting uses $5{,}000$ calibration runs and $3{,}000$ alternative runs. Figure 2 connects signal, threshold components, and power.

PBOS retains high power across the ladder, from 93.93% under weak confounding to 92.27% under strong confounding. Product clipping falls from 88.77% to 73.93%, a gap of 18.33 percentage points at the strongest setting ($2{,}768$ versus $2{,}218$ rejections out of $3{,}000$). The standardized signals are nearly tied, 4.45 for PBOS against 4.60 for product clipping at strong confounding, so the power gap must come from the thresholds.

Two terms in the threshold account for the separation. The PBOS fitting shift remains close to 0.10, while the product-clipping shift grows from 3.401 to 7.196. The concentration term is also smaller for PBOS, 14.02 compared with 27.63, as predicted by the conditional span calculation. The resulting strong-confounding thresholds are 14.43 and 35.15. On their respective null-standard-deviation scales, product clipping pays a higher threshold despite its slightly larger signal separation.

Tightening the concentration term tests this explanation more directly. In a separate paired batch, we replace both worst-case spans with their actual conditional spans, using the same records and noise. The strong-confounding power gap is 11.37 percentage points after this refinement, compared with 18.97 before it. Label-span concentration accounts for part of the difference; the remaining gain survives calibration with the sharper spans. Appendix G.10 reports the paired thresholds, rejection margins, and direct nuisance perturbations.

5.3 Audit resolution and privacy budgets

Independent primary validation yields $65/20{,}000$ null rejections, or 0.325%, against the target upper bound of 5%. At $\theta = 0.36$, all $10{,}000$ alternative runs reject. The independent effect curve also resolves smaller effects, reaching 92.93% power at $\theta = 0.20$. Its $3{,}000$ runs are distinct from the moderate-confounding ladder cell, whose estimate is 93.77%.

The second encoder follows the same progression. EmbeddingGemma reaches 88.77% power at $\theta = 0.20$ and rejects all $10{,}000$ runs at $\theta = 0.36$ under its own calibrated threshold of 14.45; its independent null validation gives $63/20{,}000$ rejections.

Under the tested privacy schedule, PBOS power rises from 72.20% at $\varepsilon = 1$ through 88.57% at $\varepsilon = 2$ to 93.77% at $\varepsilon = 4$, and the $\varepsilon = 8$ estimate of 93.17% has an interval overlapping the one at 4: the schedule's knee lies at the nominal budget. Full curves, budgets, and thresholds are provided in Appendix G.

Figure 3 connects the calibration benefit to sample requirements in an additional actual-span study with $1{,}000$ calibration pairs and 400 validation pairs per cell. Under strong confounding and $\theta = 0.20$, PBOS first exceeds 90% estimated power at $16{,}000$ records per split, compared with $24{,}000$ for product clipping. These are crossings on the tested grid; the two panels show the progression at both effect strengths. Appendix G.12 gives the protocol and complementary clipping, effect-confounding, and calibration-drift figures.

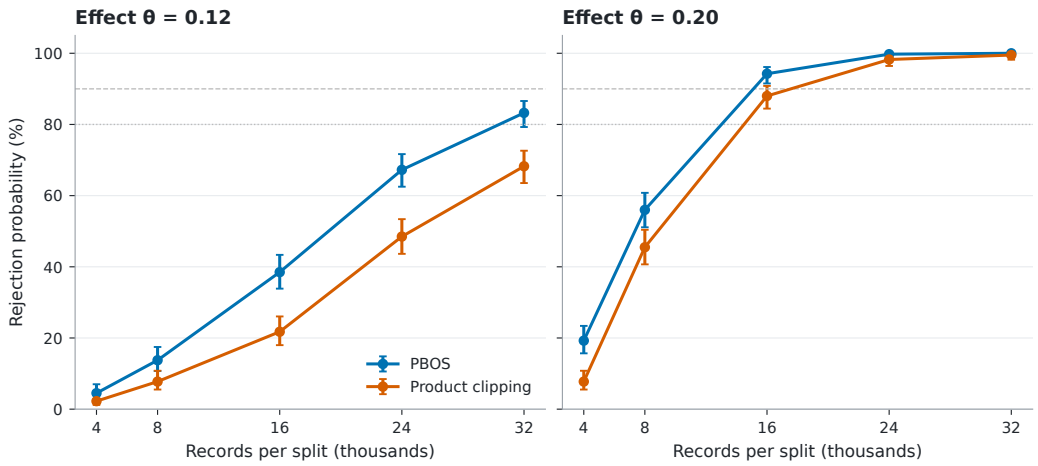


Figure 3: Sample efficiency under strong confounding. Each point uses 400 paired validation runs and pointwise Wilson 95% intervals after 1,000 calibration pairs. The horizontal axis gives records per split; total training and evaluation slots are twice this number. Horizontal lines mark 80% and 90% power. This synthetic study uses actual-span diagnostic thresholds.

5.4 Directional interpretation and additional settings

The alignment experiment identifies the component measured by the audit. As the target becomes predictable from nuisance context, PBOS rejection moves from 93.77% to 2.17%, while the unresidualized reference retains 86.60% rejection at the largest alignment (Figure 8). This matches Equation (12): PBOS measures the directional variation remaining after context is removed.

The controls further distinguish the queried direction. Nuisance-only, descriptor-permuted, and norm-matched random queries have rejection rates below 1%. A target-correlated hard control retains 19.27% power. Reversing the query sign retains 87.87% two-sided power. Together, these results connect detection to the chosen direction and the component left after removing observed context.

Source-stratified cluster resampling gives 92.50% power at $\theta = 0.20$; leave-one-source-out powers range from 88.67% to 98.57%. The strong-effect result persists in these settings. With an omitted nonlinear nuisance interaction, the methods attain comparable power: 83.77% for PBOS and 84.20% for product clipping, with overlapping intervals. The confounding ladder identifies the setting where preserving nuisance cancellation brings the largest gain.

We also apply the frozen query to Civil Comments toxicity labels (Borkan et al., 2019), using disjoint public-pilot, training, and held-out evaluation sets. On 12,000 evaluation records, the PBOS mean is -0.003204, with record-bootstrap 95% interval $[-0.003572, -0.002847]$: a nonzero held-out association for the fitted score with naturally annotated labels. The preceding calibrated tests use specified null laws; Appendix H reports all three score summaries and the data protocol.

6 Discussion

The scalar-preservation theorem and the residual-product example isolate the decisive step: differentiation after bounding reweights the nuisance contributions. PBOS preserves the cancellation by centering a bounded multiplier before combining it with the binary label residual. In the exact model this gives the product-rate condition; in the fitted audits it gives a smaller shift to calibrate.

The two calibration routes use different information: null simulation measures the shift of the complete fitting procedure, and nuisance-error bounds control it through the product remainder and the bounded-root defect. The representation experiments show why this matters; scores with similar standardized signal can have substantially different power once their fitted null shifts enter the threshold. A practitioner choosing a bound for a private score should therefore treat placement as part of the design, not as a post-processing step that orthogonalization has already taken care of.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript polishing, and code development. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

The study analyzes fixed embeddings of public English text, synthetic labels, and existing toxicity annotations on a public Civil Comments subset. Toxicity labels reflect annotator judgments and are used only for an aggregate descriptive audit.

REPRODUCIBILITY STATEMENT

The appendices provide complete proofs, nuisance objectives, the sampling law, privacy and calibration assumptions, and expanded experimental results. The accompanying source package contains the figure data and scripts and the recorded calibration and validation summaries. The numerical checks reproduce reported thresholds, rejection rates, and power-certificate aggregation from the supplied records.

REFERENCES

Daniel G. Alabi and Salil P. Vadhan. Differentially private hypothesis testing for linear regression. *Journal of Machine Learning Research*, 24(361):1–50, 2023. URL https://www.jmlr.org/papers/v24/23-0045.html.

Amir Asiaee and Kaveh Aryan. Workload-preserving differentially private synthetic data for causal inference via maximum-entropy calibration. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 133–160. PMLR, 2026. URL https://proceedings.mlr.press/v337/asiaee26a.html.

Marco Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021. doi: 10.1080/01621459.2019.1700130. URL https://doi.org/10.1080/01621459.2019.1700130.

Marco Avella-Medina. Seconder of the vote of thanks to Dong et al. and contribution to the discussion of "Gaussian Differential Privacy". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):39–41, 2022. doi: 10.1111/rssb.12456. URL https://doi.org/10.1111/rssb.12456.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. ACM, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Clément L. Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 310–321. ACM, 2019. doi: 10.1145/3313276.3316336. URL https://doi.org/10.1145/3313276.3316336.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15676–15688, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract.html.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097. URL https://academic.oup.com/ectj/article/21/1/C1/5056401.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022. doi: 10.3982/ECTA16294. URL https://doi.org/10.3982/ECTA16294.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, pp. 371–380. ACM, 2009. doi: 10.1145/1536414.1536466. URL https://doi.org/10.1145/1536414.1536466.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042. URL https://www.nowpublishers.com/article/Details/TCS-042.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023. doi: 10.1214/23-AOS2258. URL https://doi.org/10.1214/23-AOS2258.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Iden Kalemaj, Shiva Kasiviswanathan, and Aaditya Ramdas. Differentially private conditional independence testing. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3700–3708. PMLR, 2024. URL https://proceedings.mlr.press/v238/kalemaj24a.html.

Gautam Kamath, Argyris Mouzakis, Matthew Regehr, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A bias-accuracy-privacy trilemma for statistical estimation. *Journal of the American Statistical Association*, 120(552):2338–2349, 2025. doi: 10.1080/01621459.2024.2443275. URL https://doi.org/10.1080/01621459.2024.2443275.

Zeki Kazan, Kaiyan Shi, Adam Groce, and Andrew P. Bray. The test of tests: A framework for differentially private hypothesis testing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16131–16151. PMLR, 2023. URL https://proceedings.mlr.press/v202/kazan23a.html.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17994–18008, 2022. doi: 10.52202/068431-1308. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/725f5e8036cc08adeba4a7c3bcbc6f2c-Abstract-Conference.html.

National Center for Biotechnology Information. PMC open access subset. PubMed Central, 2026. URL https://pmc.ncbi.nlm.nih.gov/tools/openftlist/. Data resource; accessed 2026-09-06.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022. URL https://www.jmlr.org/papers/v23/21-1328.html.

Maresa Schröder, Valentyn Melnychuk, and Stefan Feuerriegel. Differentially private learners for heterogeneous treatment effects. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6baec7c4ba0a8734ccbd528a8090cb1f-Abstract-Conference.html.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-AOS1857. URL https://doi.org/10.1214/19-AOS1857.

Stack Exchange, Inc. Stack exchange data dump. Internet Archive, 2014. URL https://archive.org/details/stackexchange. Data collection first posted 2014-01-21; accessed 2026-09-06.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, et al. EmbeddingGemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Zhanyu Wang, Guang Cheng, and Jordan Awan. Differentially private bootstrap: New privacy analysis and inference strategies. *Journal of Machine Learning Research*, 26(257):1–57, 2025. URL https://www.jmlr.org/papers/v26/23-0514.html.

Wikimedia Foundation. English Wikipedia data dumps. Wikimedia Data Dumps, 2026. URL https://meta.wikimedia.org/wiki/Research:Data. Data portal; accessed 2026-09-06.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://doi.org/10.1080/01621459.1927.10502953.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

The proofs follow the dependence of inference on score construction. Appendices B and C identify the nuisance moments changed by bounding and the moments PBOS centers. Appendix D turns that difference into a separation in admissible nuisance rates, then derives calibration thresholds from error envelopes. Appendix E calibrates the complete fitting error. The corresponding experiments appear in Appendices F, G, and H; Appendix I carries the fitted score through a lattice release.

A Additional background and privacy

A.1 Audit target and data access

Variables: For each evaluation unit, $Y_i \in \{0,1\}$ is a binary target, $Z_i \in \mathbb{R}$ is a fixed directional representation score, and N_i is the prespecified nuisance information. The population anchors are

$$q_0(N) = \mathbb{P}(Y = 1 \mid N), \qquad A_0 = Z - g_0(N).$$

The audit direction is fixed before confirmatory evaluation. The directional population moment for an odd bounded multiplier d is

$$\mu_d = \mathbb{E}[\{Y - q_0(N)\}d(A_0)].$$

Sample splitting: The training set I_1 and evaluation set I_2 are disjoint at the privacy adjacency unit. The public transcript $\mathcal{F}_{\mathrm{pub}}$ contains public features, the prespecified mechanism configuration, and public constants. The pre-label transcript $\mathcal{F}_{\mathrm{pre}}$ additionally contains the realized private nuisance outputs, all fixed coefficients used by the evaluation statistic, and the public threshold, but not the fresh evaluation labels. Conditional statements about the evaluation statistic are made given $\mathcal{F}_{\mathrm{pre}}$.

Released information: The audit returns its prespecified public identifiers, noisy statistic, and threshold decision. Residuals and fitted coefficients remain internal. This choice also excludes optimization traces and unnoised sums; the privacy proofs concern this observable channel.

A.2 What the directional moment measures

Proposition A.1 (Conditional-independence bridge)**.** *Let $A_0 = Z - g_0(N)$ be measurable with respect to $\sigma(Z, N)$, and suppose $q_0(N) = \mathbb{P}(Y = 1 \mid N)$ almost surely. If $Y \perp Z \mid N$ and $\mathbb{E}|d(A_0)| < \infty$, then*

$$\mathbb{E}[\{Y - q_0(N)\}d(A_0)] = 0.$$

Proof. Conditional independence and measurability give

$$\mathbb{E}[Y - q_0(N) \mid Z, N] = \mathbb{E}[Y - q_0(N) \mid N] = 0.$$

Multiplying by the integrable $\sigma(Z, N)$-measurable variable $d(A_0)$ and applying iterated expectation proves the claim. □

Proposition A.2 (A zero directional moment need not imply conditional independence)**.** *Suppose the conditional law of A_0 is symmetric about zero and d is odd. Let h be a bounded even function with $\mathbb{E}[h(A_0) \mid N] = 0$ and with $\mathrm{Var}\{h(A_0) \mid N\} > 0$ on an event of positive probability. Choose $\epsilon \neq 0$ such that $|\epsilon|\|h(A_0)\|_\infty \leq 1/2$, and let*

$$\mathbb{P}(Y = 1 \mid Z, N) = \frac{1}{2} + \epsilon h(A_0).$$

Then $Y \not\perp Z \mid N$, but

$$\mathbb{E}\left[\left(Y - \frac{1}{2}\right) d(A_0) \,\middle|\, N\right] = 0.$$

Proof. The conditional variance condition makes the label probability depend on Z through A_0 on an event of positive probability, so conditional independence fails. Boundedness and the restriction on ϵ make the displayed probability valid. Moreover,

$$\mathbb{E}\left[\left(Y - \frac{1}{2}\right) d(A_0) \,\middle|\, N\right] = \epsilon\mathbb{E}[h(A_0)d(A_0) \mid N].$$

The product of the even function h and odd function d is odd. Symmetry of the conditional residual law makes its expectation zero. □

The two propositions explain the interpretation of the audit. Conditional independence implies a zero score mean when the prediction anchor is correct. The reverse implication fails even for bounded binary models: an even dependence on the residual is invisible to an odd query. Thus, a single PBOS audit answers a directional question.

A.3 Privacy of the nominal component mechanisms

Proposition A.3 (Disjoint train and evaluation composition)**.** *Suppose the complete training mechanism is $(\varepsilon_{\mathrm{tr}}, \delta_{\mathrm{tr}})$-DP on I_1, and for every fixed training output the evaluation mechanism is $(\varepsilon_{\mathrm{ev}}, \delta_{\mathrm{ev}})$-DP on the disjoint set I_2. Assume evaluation accesses training records only through that private training output, and the complete observable output is a data-independent post-processing of these outputs and public inputs. Under replace-one adjacency at the disjoint slot level, with split membership fixed independently of private record values, the release is*

$$(\max\{\varepsilon_{\mathrm{tr}}, \varepsilon_{\mathrm{ev}}\}, \max\{\delta_{\mathrm{tr}}, \delta_{\mathrm{ev}}\})\text{-}DP.$$

Proof. One adjacent replacement affects at most one of the disjoint partitions. For a training replacement, the evaluation computation with fixed evaluation records is post-processing of the private training output. For an evaluation replacement, the training-output distribution is unchanged, and the conditional evaluation guarantee holds uniformly over it. Taking the maximum parameters across these two cases and applying the fixed release projection gives the result (Dwork & Roth, 2014). □

Primary parameters: Both nuisance fits use the training split, so their component budgets add: $(\varepsilon_\eta, \varepsilon_\omega) = (3, 1)$ and $(\delta_\eta, \delta_\omega) = (5 \times 10^{-7}, 5 \times 10^{-7})$ give $(\varepsilon_{\mathrm{tr}}, \delta_{\mathrm{tr}}) = (4, 10^{-6})$. The evaluation mechanism uses $(\varepsilon_{\mathrm{ev}}, \delta_{\mathrm{ev}}) = (4, 10^{-6})$. Under the output-channel premise of Proposition A.3, parallel composition gives $(4, 10^{-6})$-DP.

A.4 Noise calibration in the experiments

The nominal fitting procedure perturbs two ridge solutions. Treating both solutions as private outputs gives the following component accounting; the actual audit retains them internally. Under the declared feature bound $\|W_3\|_2 \le l_W = 2$, replace-one adjacency, and the strong-convexity conditions of the two ridge objectives, the implementation uses the sensitivity bounds

$$\Delta_\eta = \frac{2l_W}{n_{\mathrm{tr}}\lambda_\eta}, \qquad \Delta_\omega = \frac{Bl_W}{2n_{\mathrm{tr}}\lambda_\omega}.$$

The evaluation release is a sum of contributions in $[-B, B]$ and has replacement sensitivity $\Delta_{\mathrm{ev}} = 2B$. For a target pair (ε, δ), the continuous analytic Gaussian mechanism selects σ so that

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \le \delta$$

(Balle & Wang, 2018). With $n_{\mathrm{tr}} = 16{,}000$, $(\lambda_\eta, \lambda_\omega) = (0.4, 0.2)$, and $B = 0.08$, the declared sensitivities are 6.25×10^{-4}, 2.5×10^{-5}, and 0.16. The corresponding minimum continuous Gaussian scales are approximately 9.923×10^{-4} for η, 1.091×10^{-4} for ω, and 0.1910 for evaluation. The implementation uses these calibrated training scales and the larger evaluation scale $\sigma_{\mathrm{ev}} = 1$. Thus the component budgets used in Proposition A.3 follow from the continuous Gaussian mechanisms under the declared sensitivity conditions. Section I states the separate finite-precision sampler requirement.

B Proofs for post-hoc bounding

Scalar bounding changes the weight assigned to each nuisance derivative. The two-state construction below shows how that weight removes an exact cancellation while preserving the null center. Requiring cancellation for every centered score path then forces the transformation to be affine.

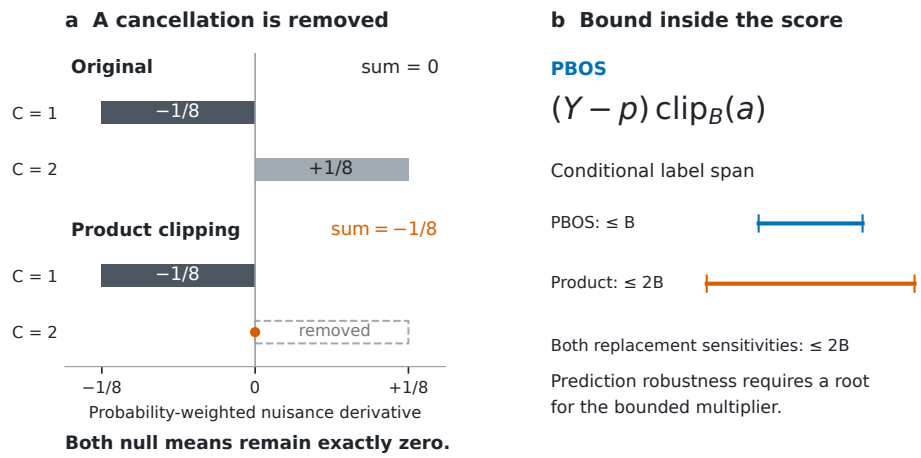


Figure 4: **Clipping removes a cancellation in the two-state example.** Left: probability-weighted nuisance derivatives before and after clipping, with both null means zero. Right: worst-case conditional label spans for the two placements. This is the tangent example below; the residual-product order-gap model appears in Appendix D.

B.1 Differentiating the transformed score

Assumption B.4 (Path regularity)**.** The nuisance path $t \mapsto \psi_t$ is differentiable at $t = 0$ almost surely, with $\dot{\psi} = \partial_t \psi_t|_{t=0}$. Its difference quotients are dominated in L^1 in a neighborhood of zero. The transform b is Lipschitz and differentiable outside a null set that has zero probability under ψ_0.

Proposition B.5 (Derivative after post-hoc bounding)**.** *Under Assumption B.4,*

$$\partial_t \mathbb{E}[b(\psi_t)]|_{t=0} = \mathbb{E}[b'(\psi_0)\dot{\psi}].$$

Consequently, $\mathbb{E}\dot{\psi} = 0$ *does not in general imply a zero transformed derivative.*

Proof. At every sample point where both derivatives exist, the ordinary chain rule gives

$$\lim_{t\to 0} \frac{b(\psi_t) - b(\psi_0)}{t} = b'(\psi_0)\dot{\psi}.$$

If L_b is a Lipschitz constant, the absolute difference quotient is bounded by $L_b|\psi_t - \psi_0|/|t|$. The assumed L^1 envelope permits dominated convergence. The second statement follows because $b'(\psi_0)$ need not be constant or uncorrelated with $\dot{\psi}$. □

B.2 A centered score that loses orthogonality

Let a public state C be uniform on $\{1, 2\}$. Set

$$(w, a) = \begin{cases} (1, 1), & C = 1, \\ (-1/2, 2), & C = 2, \end{cases} \qquad Y \mid C \sim \mathrm{Bernoulli}(1/2).$$

Define $p_t(C) = \Lambda(tw)$ and $\psi_t = (Y - p_t)a$. At zero,

$$\dot{\psi} = -\Lambda(0)\{1 - \Lambda(0)\}wa = -\frac{1}{4}wa.$$

Thus $\dot{\psi} = -1/4$ on $C = 1$ and $\dot{\psi} = 1/4$ on $C = 2$, so $\mathbb{E}\dot{\psi} = 0$. Also, conditional on C, ψ_0 is symmetric about zero: it takes values $\pm 1/2$ when $C = 1$ and ± 1 when $C = 2$. Hence $\mathbb{E}\psi_0 = 0$.

For $B = 3/4$, the clipped values remain symmetric, so $\mathbb{E}\,\mathrm{clip}_{3/4}(\psi_0) = 0$. Only the $C = 1$ observations lie in the linear region. Proposition B.5 therefore yields

$$\partial_t \mathbb{E}\,\mathrm{clip}_{3/4}(\psi_t)\Big|_{t=0} = \frac{1}{2}\left(-\frac{1}{4}\right) = -\frac{1}{8}.$$

For $b(u) = \tanh(u)$,

$$\partial_t \mathbb{E}\tanh(\psi_t)|_{t=0} = \frac{1}{8}\{\mathrm{sech}^2(1) - \mathrm{sech}^2(1/2)\} \approx -0.0458.$$

Both transforms produce a nonzero derivative. The same change occurs for a smooth transform: the source is the unequal weighting of nuisance derivatives.

Post-hoc clipping can also move a null center. If $G = 2$ with probability $1/3$ and $G = -1$ with probability $2/3$, then $\mathbb{E}G = 0$, whereas $\mathbb{E}\,\mathrm{clip}_1(G) = -1/3$. In the Bernoulli construction, centering survives. It therefore isolates the change in nuisance sensitivity from a change in the null mean.

B.3 When scalar bounding preserves cancellation

Corollary B.6 (Exact clipped-tangent identity)**.** *Under Assumption B.4, if $\mathbb{P}(|\psi_0| = B) = 0$, then*

$$\partial_t \mathbb{E}\,\mathrm{clip}_B(\psi_t)|_{t=0} = \mathbb{E}[\dot{\psi}\mathbf{1}\{|\psi_0| < B\}].$$

If $\mathbb{E}\dot{\psi} = 0$, the derivative equals

$$-\mathbb{E}[\dot{\psi}\mathbf{1}\{|\psi_0| \geq B\}].$$

Hard clipping preserves orthogonality if and only if the clipped set carries no net nuisance tangent.

Proof. Away from $\{-B, B\}$, the derivative of $u \mapsto \mathrm{clip}_B(u)$ equals $\mathbf{1}\{|u| < B\}$. Proposition B.5 proves the first equality. Splitting $\mathbb{E}\dot{\psi}$ over the clipped and unclipped regions gives the second. □

Corollary B.7 (No universal nonconstant bounded scalar repair)**.** *Let $b : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz. Suppose that $\mathbb{E}[b'(S)H] = 0$ for every finitely supported pair (S, H) satisfying $\mathbb{E}S = \mathbb{E}H = 0$ for which b is differentiable at each support value of S. Then b is affine. If b is globally bounded, it is constant.*

Proof. Let D be the differentiability set of b, which has full Lebesgue measure. Fix $x, y \in D$. If $x + y = 0$, place equal mass on $(S, H) = (x, 1)$ and $(y, -1)$; the hypothesis gives $b'(x) = b'(y)$. Otherwise choose $z \in D$ with sign opposite to $x + y$ and set

$$a = \frac{-z}{x + y - 2z} \in (0, 1/2).$$

Assign masses $a, a, 1 - 2a$ to $(x, 1), (y, -1), (z, 0)$. Then $\mathbb{E}S = \mathbb{E}H = 0$, and the hypothesis gives $a\{b'(x) - b'(y)\} = 0$. Thus b' is constant almost everywhere. A Lipschitz function is absolutely continuous on every compact interval, so integrating its derivative shows that b is affine. Global boundedness forces its slope to vanish. □

The proof uses only paths $\psi_t = S + tH$ that stay centered for all t. Centered paths alone thus force the affine conclusion, and local Lipschitz regularity includes hard clipping. For $b \in C^1$, equal mass on $(x, 1), (y, -1), (-x - y, 0)$ gives $b'(x) = b'(y)$ directly. For a fixed score and nuisance path, the covariance below gives the exact preservation condition.

Proposition B.8 (Exact safety characterization)**.** *Under Assumption B.4 and $\mathbb{E}\dot{\psi} = 0$,*

$$\partial_t \mathbb{E}b(\psi_t)|_{t=0} = \mathrm{Cov}\{b'(\psi_0), \dot{\psi}\}.$$

Thus the transformed score is orthogonal if and only if $b'(\psi_0)$ and $\dot{\psi}$ are uncorrelated. Sufficient conditions include an almost surely constant $b'(\psi_0)$ or

$$\mathbb{E}[\dot{\psi} \mid b'(\psi_0)] = 0.$$

For hard clipping, the latter is $\mathbb{E}[\dot{\psi} \mid \mathbf{1}\{|\psi_0| < B\}] = 0$.

Proof. Since $\mathbb{E}\dot{\psi} = 0$, Proposition B.5 gives

$$\mathbb{E}[b'(\psi_0)\dot{\psi}] = \mathbb{E}[b'(\psi_0)\dot{\psi}] - \mathbb{E}[b'(\psi_0)]\mathbb{E}\dot{\psi},$$

which is the covariance. The first sufficient condition makes this covariance zero directly. Under the second,

$$\mathbb{E}[b'(\psi_0)\dot{\psi}] = \mathbb{E}[b'(\psi_0)\mathbb{E}\{\dot{\psi} \mid b'(\psi_0)\}] = 0.$$

□

C PROOFS FOR PBOS

C.1 WHY THE BERNOULLI RESIDUAL APPEARS

Lemma C.9 (Canonical factorization)**.** *Let $Y \sim \mathrm{Bernoulli}(p)$ with $p \in (0,1)$. If $s(\cdot; p, a)$ is centered for every (p, a), so that*

$$ps(1; p, a) + (1-p)s(0; p, a) = 0,$$

then a unique $c(p, a)$ satisfies

$$s(Y; p, a) = (Y - p)c(p, a), \qquad c(p, a) = \frac{s(1; p, a)}{1-p} = -\frac{s(0; p, a)}{p}.$$

Proof. The centering equation implies $s(1; p, a)/(1-p) = -s(0; p, a)/p$. Call the common value $c(p, a)$. Substitution for $Y = 1$ and $Y = 0$ gives the factorization. Either support point determines c, proving uniqueness. □

Centering determines the label factor and leaves the multiplier c free. PBOS uses that freedom to bound the representation residual. The Bernoulli centering identity then holds before the nuisance derivatives enter the argument.

C.2 NUISANCE AND SCIENTIFIC DERIVATIVES

Let P_θ keep the marginal law of (Z, N) fixed while changing the conditional label mean $m_\theta(Z, N) = \mathbb{E}_\theta[Y \mid Z, N]$. Assume $m_0(Z, N) = q_0(N)$. For nuisance paths $q_t = q_0 + th_q$ and $g_s = g_0 + sh_g$, define

$$\mathcal{M}(\theta, t, s) = \mathbb{E}_\theta[\{Y - q_t(N)\}d\{Z - g_s(N)\}].$$

Theorem C.10 (Generic PBOS tangent separation)**.** *Suppose h_q is integrable, the scientific derivative can pass through expectation, $q_t(N) \in [0, 1]$ along the admissible prediction paths, and d is bounded by B. Then*

$$\mathcal{M}(0, 0, s) = 0, \qquad D_s\mathcal{M}(0, 0, 0)[h_g] = 0,$$

for every admissible s and h_g. Moreover,

$$D_t\mathcal{M}(0, 0, 0)[h_q] = -\mathbb{E}\{h_q(N)d(A_0)\}, \qquad D_\theta\mathcal{M}(0, 0, 0) = \mathbb{E}\{\dot{m}_0(Z, N)d(A_0)\},$$

where $A_0 = Z - g_0(N)$ and $\dot{m}_0 = \partial_\theta m_\theta|_{\theta=0}$. Every summand is bounded by B, and replacing one record changes the sum by at most $2B$.

Proof. At the null, iterated expectation gives

$$\mathcal{M}(0, 0, s) = \mathbb{E}[\mathbb{E}\{Y - q_0(N) \mid Z, N\}d\{Z - g_s(N)\}] = 0.$$

The equality holds along the entire residual-nuisance path, so its derivative at zero is also zero. Differentiating q_t gives the prediction-nuisance derivative directly. Because the marginal law of (Z, N) is fixed along P_θ,

$$\mathcal{M}(\theta, 0, 0) = \mathbb{E}[\{m_\theta(Z, N) - q_0(N)\}d(A_0)],$$

which yields the scientific derivative. Finally, $|Y - q_t(N)| \le 1$ and $|d| \le B$ imply the contribution and replacement bounds. □

The residual-fitting identity holds along the entire nuisance path. No derivative of d enters its proof, which is why hard clipping is admissible. For prediction fitting, $\mathbb{E}[d(A_0) \mid N] = 0$ cancels every integrable prediction tangent by centering the bounded multiplier itself. This condition determines the residual center and yields the product-error bound below.

Proposition C.11 (Second-order remainder under a conditional residual root)**.** *Assume $\mathbb{E}[Y \mid Z, N] = q_0(N)$, d is L-Lipschitz, and $\mathbb{E}[d(A_0) \mid N] = 0$, where $A_0 = Z - g_0(N)$. For square-integrable nuisance errors,*

$$|\mathbb{E}[\{Y - q(N)\}d\{Z - g(N)\}]| \le L\|q - q_0\|_2\|g - g_0\|_2.$$

Proof. Iterated expectation replaces Y by $q_0(N)$. Add and subtract $d(A_0)$; the conditional root removes the term $\mathbb{E}[\{q_0 - q\}d(A_0)]$. The remaining integrand has absolute value at most $L|q - q_0||g - g_0|$. Cauchy–Schwarz gives the bound. □

Section C.3 treats finite-dimensional residualization through its own moment condition.

Corollary C.12 (Logistic residual-energy specialization)**.** *Assume d is odd and nondecreasing. Consider the logistic directional path*

$$m_\theta(Z, N) = \Lambda\{\operatorname{logit} q_0(N) + \theta Z\}, \qquad \rho_0 = q_0(N)\{1 - q_0(N)\}.$$

If the population center g_0 satisfies the weighted residualization condition

$$\mathbb{E}[\rho_0 g_0(N)d\{Z - g_0(N)\}] = 0, \tag{C.1}$$

then

$$D_\theta\mathcal{M}(0, 0, 0) = \mathbb{E}[\rho_0 A_0 d(A_0)] \ge 0.$$

The inequality is strict whenever $A_0 d(A_0) > 0$ with positive probability on a set where $\rho_0 > 0$.

Proof. The logistic path has $\dot{m}_0 = \rho_0 Z$. Substituting $Z = A_0 + g_0(N)$ into Theorem C.10 gives

$$D_\theta\mathcal{M}(0, 0, 0) = \mathbb{E}[\rho_0 A_0 d(A_0)] + \mathbb{E}[\rho_0 g_0(N)d(A_0)].$$

Equation (C.1) removes the second term. Oddness and monotonicity give $A_0 d(A_0) \ge 0$, and the stated nondegeneracy condition gives strict positivity. □

C.3 Logistic prediction and residualization

Assume $d : \mathbb{R} \to \mathbb{R}$ is odd, nondecreasing, L_d-Lipschitz, and $|d(a)| \le B$. The experiments use $d(a) = \operatorname{clip}_B(a)$ with $B = 0.08$. A smooth alternative is $d(a) = B\tanh(ca/B)$, whose derivative at zero is c.

Let $\widetilde{w}_{\eta i}$ and $\widetilde{w}_{\omega i}$ be the fixed feature vectors used for prediction and residualization. Their coefficient anchors are x_0 and y_0, respectively, and z_i is the directional score. Define

$$q_{\theta i} = \Lambda(\theta z_i + \widetilde{w}_{\eta i}^\top x_0), \qquad q_i = q_{0i}, \qquad \rho_i = q_i(1 - q_i),$$
$$a_i^0 = z_i - \widetilde{w}_{\omega i}^\top y_0, \qquad m_i = \mathbb{E}[\rho_i d(a_i^0)\widetilde{w}_{\eta i} \mid \mathcal{F}_{\mathrm{pub}}].$$

The aggregate certificate is $\|\sum_i m_i\| \le \nu_{\mathrm{orth,agg}}$. Define

$$M(\theta, x, y) = \sum_{i \in I_2} \mathbb{E}\left[\{q_{\theta i} - \Lambda(\widetilde{w}_{\eta i}^\top x)\}d(z_i - \widetilde{w}_{\omega i}^\top y) \mid \mathcal{F}_{\mathrm{pub}}\right].$$

Theorem C.13 (Signal and nuisance tangent separation)**.** *Assume differentiation can pass through the conditional expectations. Then:*

1. *$|(Y - p)d(a)| \le B$, so replacement sensitivity is at most $2B$.*
2. *$M(0, x_0, y) = 0$ for every admissible y, hence $D_y M(0, x_0, y_0) = 0$.*
3. *For every direction h,*

$$D_x M(0, x_0, y_0)[h] = -h^\top \sum_{i \in I_2} m_i.$$

This derivative is zero under aggregate exact orthogonality and has magnitude at most $\nu_{\mathrm{orth,agg}}\|h\|$ under the approximate certificate.

4. If

$$\sum_{i\in I_2} \mathbb{E}[\rho_i d(a_i^0)\widetilde{w}_{\omega i} \mid \mathcal{F}_{\text{pub}}] = 0,$$

then

$$D_\theta M(0, x_0, y_0) = \sum_{i\in I_2} \mathbb{E}[\rho_i a_i^0 d(a_i^0) \mid \mathcal{F}_{\text{pub}}] \geq 0.$$

The inequality is strict if $\rho_i > 0$ and $a_i^0 d(a_i^0) > 0$ jointly with positive conditional probability for some evaluation unit.

Proof. The first statement follows from $|Y - p| \leq 1$ and $|d| \leq B$; replacing one record changes a bounded sum by at most $2B$.

At $(\theta, x) = (0, x_0)$, the two probabilities in each summand are identical. Their difference is zero pointwise for every y, proving the second statement.

For the third statement, differentiate the logistic prediction with respect to x at x_0:

$$D_x\Lambda(\widetilde{w}_{\eta i}^\top x)[h] = \rho_i \widetilde{w}_{\eta i}^\top h.$$

The negative sign in M gives the displayed derivative. Cauchy–Schwarz and the aggregate certificate give the approximate bound.

For the target derivative,

$$D_\theta q_{\theta i}|_{\theta=0} = \rho_i z_i.$$

Since $z_i = a_i^0 + \widetilde{w}_{\omega i}^\top y_0$,

$$\begin{aligned} D_\theta M(0, x_0, y_0) &= \sum_i \mathbb{E}[\rho_i z_i d(a_i^0) \mid \mathcal{F}_{\text{pub}}] \\ &= \sum_i \mathbb{E}[\rho_i a_i^0 d(a_i^0) \mid \mathcal{F}_{\text{pub}}] + y_0^\top \sum_i \mathbb{E}[\rho_i d(a_i^0)\widetilde{w}_{\omega i} \mid \mathcal{F}_{\text{pub}}]. \end{aligned}$$

The residualization condition removes the second term. Oddness and monotonicity imply $ad(a) \geq 0$, proving nonnegativity and the stated strictness condition. □

The population derivatives identify which fitting moments control the mean shift. Appendix E measures the resulting error of the complete fitted program through calibration.

The prediction condition uses $\widetilde{w}_{\eta i}$, whereas the residual-energy condition uses $\widetilde{w}_{\omega i}$. Each centers the final bounded multiplier against the features appearing in its derivative. A span-inclusion argument is needed to transfer one condition to the other when the feature spans differ.

C.4 Comparing the two bound placements

Let $x = x_0 + \delta x$ and compare

$$\widetilde{\psi}_B = \text{clip}_B\{(Y - p)a^0\}, \qquad \phi_d = (Y - p)d(a^0).$$

Write $\Sigma_{B,i} = \{|(Y - q_i)a_i^0| < B\}$. Applying the path derivative from Appendix B to the product score gives the leading post-hoc nuisance term

$$-\delta x^\top \sum_i \mathbb{E}[\rho_i a_i^0 \mathbf{1}_{\Sigma_{B,i}} \widetilde{w}_{\eta i} \mid \mathcal{F}_{\text{pub}}] + o(\|\delta x\|).$$

A direct Taylor expansion of the PBOS prediction component gives

$$-\delta x^\top \sum_i m_i + O(\|\delta x\|^2).$$

The two expressions identify exactly what must be centered. Product clipping introduces a label-dependent indicator into the nuisance moment. PBOS instead uses the bounded residual moment specified by its fitting condition.

On the two-state witness above, replacing the product clip by $(Y - p_t)\,\mathrm{clip}_{3/4}(a)$ gives derivative $-3/64$. Thus, moving the bound alone does not ensure prediction orthogonality on an arbitrary example. The multiplier must also satisfy the prediction-centering condition in Theorem C.10 or its logistic counterpart.

For fixed p, a, PBOS has two label branches $(1-p)d(a)$ and $-p\,d(a)$, separated by $|d(a)| \le B$. The post-hoc product has branch separation at most $2B$, attained when both branches saturate. Thus the worst-case fresh-label Hoeffding variance proxy is $mB^2/4$ for PBOS and mB^2 for the comparator. Both scores still have the same full-record replacement sensitivity bound $2B$; this branch-span comparison does not halve their differential-privacy noise requirement.

D Nuisance error orders and inference

The separation in testing validity comes from the order of the null mean: product clipping leaves a prediction-error term, while PBOS retains a product of nuisance errors. We compute both means exactly, then compare them with the sampling fluctuation under inherited zero-center calibration. The remaining results quantify the same mean shift through nuisance error envelopes and the numerical fitting conditions.

D.1 An exact residual-product construction

Let $Y \sim \mathrm{Bernoulli}(1/2)$ be independent of Z, where

$$\mathbb{P}(Z = 1) = \frac{2}{3}, \qquad \mathbb{P}(Z = -2) = \frac{1}{3}.$$

The ordinary residual center is $\mathbb{E}Z = 0$. For prediction and residual-center perturbations (t, s), define

$$\begin{aligned} \psi^{t,s}_{\mathrm{raw}} &= (Y - 1/2 - t)(Z - s), \\ \psi^{t,s}_{\mathrm{post}} &= \mathrm{clip}_{3/4}\{(Y - 1/2 - t)(Z - s)\}, \\ \psi^{t,s}_{\mathrm{PB}} &= (Y - 1/2 - t)\,\mathrm{clip}_{3/4}(Z - 5/8 - s). \end{aligned} \tag{D.2}$$

The prediction is $1/2 + t$ in each score. The residual perturbation s is measured from the ordinary center for the first two scores and from the bounded-multiplier center for PBOS.

Theorem D.14 (Exact nuisance error orders)**.** *For $|t| \le 1/16$ and $|s| \le 1/8$, the scores in* (D.2) *satisfy*

$$\mathbb{E}\psi^{t,s}_{\mathrm{raw}} = ts, \qquad \mathbb{E}\psi^{t,s}_{\mathrm{post}} = -\frac{2}{3}t(1-s), \qquad \mathbb{E}\psi^{t,s}_{\mathrm{PB}} = \frac{2}{3}ts. \tag{D.3}$$

All three are centered at $(0, 0)$. The raw score and PBOS have zero derivatives in both nuisance coordinates there, whereas the prediction derivative of the post-hoc score equals $-2/3$.

Proof. Independence gives $\mathbb{E}\psi^{t,s}_{\mathrm{raw}} = (-t)(-s) = ts$. For $Z = 1$, both label branches of the raw product obey

$$|(Y - 1/2 - t)(1 - s)| \le \frac{9}{16}\frac{9}{8} = \frac{81}{128} < \frac{3}{4}.$$

For $Z = -2$, their absolute values are at least

$$|(Y - 1/2 - t)(-2 - s)| \ge \frac{7}{16}\frac{15}{8} = \frac{105}{128} > \frac{3}{4}.$$

Thus the $Z = 1$ branches stay linear and have conditional mean $-t(1 - s)$. The $Z = -2$ branches saturate to $-3/4$ for $Y = 1$ and $3/4$ for $Y = 0$; their conditional mean is zero. Weighting by the probabilities of Z proves the second identity.

At $g_H = 5/8$, the bounded multiplier takes values $3/8$ and $-3/4$, whose expectation is zero. Over the stated neighborhood, its perturbed values are exactly

$$\mathrm{clip}_{3/4}(1 - 5/8 - s) = 3/8 - s, \qquad \mathrm{clip}_{3/4}(-2 - 5/8 - s) = -3/4.$$

Their expectation is $-(2/3)s$. Independence from Y gives the third identity. Differentiation of these polynomials proves the derivative claims. □

All three scores have the same zero null mean at the anchor. What separates them is the remainder away from it. Centering the bounded multiplier preserves the product ts that controls the original residual score; clipping the product leaves a term linear in the prediction error.

The variance calculation fixes the scale against which these means are compared. In the same neighborhood, let $v_j(t,s) = \mathrm{Var}(\psi_j^{t,s})$. Direct calculation gives

$$v_{\mathrm{post}}(t,s) = \frac{1}{6}(1-s)^2 + \frac{3}{16} + \frac{2}{9}t^2(1-s)^2, \tag{D.4}$$

$$v_{\mathrm{PB}}(t,s) = \left(\frac{1}{4}+t^2\right)\left(\frac{9}{32} - \frac{1}{2}s + \frac{2}{3}s^2\right) - \frac{4}{9}t^2s^2. \tag{D.5}$$

For the post-hoc score, the second moment is $(2/3)(1/4+t^2)(1-s)^2 + 3/16$, and subtracting its squared mean yields (D.4). For PBOS, independence factors the second moment into $\mathbb{E}(Y-1/2-t)^2 = 1/4+t^2$ and

$$\mathbb{E}\{\mathrm{clip}_{3/4}(Z-5/8-s)^2\} = \frac{2}{3}(3/8-s)^2 + \frac{1}{3}(3/4)^2 = \frac{9}{32} - \frac{1}{2}s + \frac{2}{3}s^2.$$

In particular,

$$v_{\mathrm{post}}(0,0) = \frac{17}{48}, \qquad v_{\mathrm{PB}}(0,0) = \frac{9}{128}. \tag{D.6}$$

D.2 A NUISANCE-RATE SEPARATION

Theorem D.15 (First-order validity under the same nuisance rates)**.** *For each m, draw m independent observations from the law above and evaluate the scores at deterministic perturbations $t_m, s_m \to 0$. Let*

$$S_{m,j} = \sum_{i=1}^{m} \psi_{j,i}^{t_m,s_m}, \qquad T_{m,j} = \frac{S_{m,j} + W_{m,j}}{\sqrt{m\widehat{v}_{m,j}}}, \qquad j \in \{\mathrm{post}, \mathrm{PB}\}.$$

Assume $W_{m,j} = o_P(\sqrt{m})$, with release noise independent of the evaluation sample, and $\widehat{v}_{m,j} \to_P v_j(0,0)$. For $0<\alpha<1$, let $z_{1-\alpha/2}$ be the standard normal quantile. If

$$\sqrt{m}\,|t_m| \longrightarrow \infty, \qquad \sqrt{m}\,|t_m s_m| \longrightarrow 0, \tag{D.7}$$

then the inherited zero-center tests satisfy

$$\mathbb{P}\{|T_{m,\mathrm{post}}| > z_{1-\alpha/2}\} \longrightarrow 1, \qquad \mathbb{P}\{|T_{m,\mathrm{PB}}| > z_{1-\alpha/2}\} \longrightarrow \alpha. \tag{D.8}$$

Proof. Write $\mu_{m,j} = \mathbb{E}\psi_j^{t_m,s_m}$ and $v_{m,j} = v_j(t_m,s_m)$. The perturbations eventually enter the neighborhood of Theorem D.14. The score contributions there are uniformly bounded, and (D.6) gives $v_{m,j} \to v_j(0,0) > 0$. Thus, for every $\epsilon > 0$, the centered contributions have absolute value below $\epsilon\sqrt{mv_{m,j}}$ for all sufficiently large m. This verifies the Lindeberg condition and proves

$$\frac{S_{m,j} - m\mu_{m,j}}{\sqrt{mv_{m,j}}} \;\Rightarrow\; N(0,1).$$

Adding $W_{m,j}$ and using consistency of $\widehat{v}_{m,j}$ gives

$$R_{m,j} := \frac{S_{m,j} - m\mu_{m,j} + W_{m,j}}{\sqrt{m\widehat{v}_{m,j}}} \;\Rightarrow\; N(0,1).$$

For PBOS, $\sqrt{m}\mu_{m,\mathrm{PB}} = (2/3)\sqrt{m}\,t_m s_m \to 0$, so $T_{m,\mathrm{PB}} \Rightarrow N(0,1)$. Continuity of the normal distribution at the two critical values gives the asserted limit α.

For the post-hoc score,

$$\frac{\sqrt{m}\,|\mu_{m,\mathrm{post}}|}{\sqrt{\widehat{v}_{m,\mathrm{post}}}} = \frac{(2/3)\sqrt{m}\,|t_m|\,|1-s_m|}{\sqrt{\widehat{v}_{m,\mathrm{post}}}} \longrightarrow_P \infty.$$

Since $R_{m,\mathrm{post}} = O_P(1)$, the triangle inequality implies $|T_{m,\mathrm{post}}| \to_P \infty$. This proves the first limit even if the signs of t_m vary with m. □

Corollary D.16 (Rate region)**.** *Let $t_m = c_q m^{-\gamma_q}$ and $s_m = c_g m^{-\gamma_g}$, where $c_q, c_g \neq 0$ and $\gamma_q, \gamma_g > 0$. The conclusion of Theorem D.15 holds whenever*

$$\gamma_q < \frac{1}{2}, \qquad \gamma_q + \gamma_g > \frac{1}{2}.$$

For common rates $\gamma_q = \gamma_g = \gamma$, this region is $1/4 < \gamma < 1/2$.

Proof. Substitute the power-law sequences into (D.7): the two magnitudes are $|c_q| m^{1/2-\gamma_q}$ and $|c_q c_g| m^{1/2-\gamma_q-\gamma_g}$. □

The fixed normal critical value makes the nuisance remainder decisive. Full matched-null calibration includes the fitted score's null shift in its reference distribution, so the experiment examines this calibration cost in its second panel. A denominator $\sqrt{m\widehat{v}_{m,j} + \sigma^2}$ with fixed release-noise variance σ^2 has the same asymptotic limits, since $\sigma^2/m \to 0$. Even at the exact ordinary residual center, $s = 0$, the post-hoc score retains its diverging shift under $\sqrt{m}|t_m| \to \infty$.

D.3 Finite-sample calibration from error envelopes

Let $\mathcal{F}$ contain the training sample and all randomness used to fit the nuisance functions. Conditional on $\mathcal{F}$, the evaluation observations $(Y_i, Z_i, N_i)_{i=1}^m$ are independent with a common fixed population law, where $Y_i \in \{0, 1\}$ and

$$\mathbb{E}[Y \mid Z, N] = q_0(N).$$

The fitted functions $\widehat{q}$ and $\widehat{g}$ are $\mathcal{F}$-measurable, and $\widehat{q}(N) \in [0, 1]$ almost surely. Let d be L-Lipschitz with $|d| \leq B$, where $B > 0$. For a reference residual center g_0, define

$$r_0(N) = \mathbb{E}[d\{Z - g_0(N)\} \mid N], \qquad X_i = \{Y_i - \widehat{q}(N_i)\} d\{Z_i - \widehat{g}(N_i)\}.$$

All L_2 norms below integrate over a fresh population draw of N, holding the fitted functions fixed.

Theorem D.17 (Size control from nuisance and root envelopes)**.** *Suppose there are fixed nonnegative bounds a, b, c, e, σ and an $\mathcal{F}$-measurable event $\mathcal{E}$, with $\mathbb{P}(\mathcal{E}) \geq 1 - \beta$, such that on $\mathcal{E}$,*

$$\|\widehat{q} - q_0\|_2 \leq a, \qquad \|\widehat{g} - g_0\|_2 \leq b, \qquad \|r_0\|_2 \leq c. \tag{D.9}$$

Conditional on $\mathcal{F}$, assume that the release noise W is independent of the evaluation sample and obeys

$$\mathbb{E}[e^{uW} \mid \mathcal{F}] \leq e^{\sigma^2 u^2/2} \quad \text{for all } u \in \mathbb{R}. \tag{D.10}$$

Let the implemented statistic be $T_{\mathrm{impl}} = \sum_i X_i + W + E_{\mathrm{num}}$, where $|E_{\mathrm{num}}| \leq e$ almost surely. Set

$$0 \leq \beta < \alpha < 1, \qquad x = \log \frac{2}{\alpha - \beta}, \qquad \Delta = Lab + ac.$$

Either of the following fixed thresholds gives $\mathbb{P}(|T_{\mathrm{impl}}| > q) \leq \alpha$:

$$q_H = m\Delta + e + \sqrt{2(mB^2 + \sigma^2)x}, \tag{D.11}$$

$$q_B = m\Delta + e + \sqrt{2\{mB^2(1/4 + a^2) + \sigma^2\}x} + \frac{2B}{3}x. \tag{D.12}$$

Proof. Fix a training realization in $\mathcal{E}$, and write $\mu = \mathbb{E}[X_i \mid \mathcal{F}]$. Conditional iterated expectation and the null mean give

$$\begin{aligned}\mu &= \mathbb{E}[\{q_0(N) - \widehat{q}(N)\} d\{Z - \widehat{g}(N)\} \mid \mathcal{F}] \\ &= \mathbb{E}[\{q_0 - \widehat{q}\} r_0 \mid \mathcal{F}] \\ &\quad + \mathbb{E}[\{q_0 - \widehat{q}\}\{d(Z - \widehat{g}) - d(Z - g_0)\} \mid \mathcal{F}].\end{aligned}$$

Cauchy–Schwarz and Lipschitz continuity therefore imply

$$|\mu| \leq ac + Lab = \Delta. \tag{D.13}$$

The assumptions $Y \in \{0, 1\}$, $\widehat{q} \in [0, 1]$, and $|d| \leq B$ give $X_i \in [-B, B]$. Hoeffding's lemma yields

$$\mathbb{E}[\exp\{u(X_i - \mu)\} \mid \mathcal{F}] \leq \exp(u^2 B^2/2).$$

Independence of the evaluation draws and (D.10) imply that $V = \sum_i (X_i - \mu) + W$ has conditional moment generating function at most $\exp\{u^2(mB^2 + \sigma^2)/2\}$. Chernoff's bound in each tail gives

$$\mathbb{P}\{|V| > \sqrt{2(mB^2 + \sigma^2)x} \mid \mathcal{F}\} \le 2e^{-x} = \alpha - \beta.$$

Together with (D.13) and $|E_{\mathrm{num}}| \le e$, this proves the conditional bound for q_H.

For the Bernstein threshold, put $D = d(Z - \widehat{g}(N))$ and decompose

$$X = A + H, \qquad A = (Y - q_0(N))D, \qquad H = (q_0(N) - \widehat{q}(N))D.$$

Under the null, $\mathbb{E}[A \mid Z, N, \mathcal{F}] = 0$. Hence $\mathbb{E}[A \mid \mathcal{F}] = 0$ and $\mathbb{E}[AH \mid \mathcal{F}] = 0$, and

$$\begin{aligned}\mathrm{Var}(X \mid \mathcal{F}) &= \mathbb{E}[A^2 \mid \mathcal{F}] + \mathrm{Var}(H \mid \mathcal{F}) \\ &\le B^2 \mathbb{E}[q_0(N)\{1 - q_0(N)\} \mid \mathcal{F}] + B^2 \|\widehat{q} - q_0\|_2^2 \\ &\le B^2(1/4 + a^2).\end{aligned}$$

Also $|X - \mu| \le 2B$. The power-series bound for a centered random variable U with $|U| \le 2B$ gives, for $|u| < 3/(2B)$,

$$\log \mathbb{E}[e^{uU} \mid \mathcal{F}] \le \frac{u^2 \mathbb{E}[U^2 \mid \mathcal{F}]}{2(1 - (2B/3)|u|)}.$$

Indeed, $\mathbb{E}|U|^k \le (2B)^{k-2}\mathbb{E}U^2$ and $k! \ge 2 \cdot 3^{k-2}$ for $k \ge 2$ bound the exponential series by a geometric series, followed by $\log(1 + v) \le v$. Combining the score and noise moment generating functions yields

$$\log \mathbb{E}[e^{uV} \mid \mathcal{F}] \le \frac{u^2 v}{2(1 - k|u|)}, \qquad v = mB^2(1/4 + a^2) + \sigma^2, \quad k = 2B/3.$$

The corresponding Chernoff bound is

$$\mathbb{P}\{|V| > \sqrt{2vx} + kx \mid \mathcal{F}\} \le 2e^{-x}.$$

For example, for $v > 0$ use $u = \sqrt{2x/v}/\{1 + k\sqrt{2x/v}\}$ in the positive tail and its negative in the other tail. This proves the conditional bound for q_B; the degenerate case follows directly or by continuity. Finally, for either threshold, integrate over $\mathcal{E}$ and add $\mathbb{P}(\mathcal{E}^c) \le \beta$ to obtain the unconditional level α. □

The bounds in (D.9) are the model-specific inputs to these thresholds. They control the conditional mean; boundedness controls the remaining fluctuation. Conditioning on the fitted functions accommodates randomized fitting, with conditional noise independence and the moment bound supplying the final concentration step. Numerical error may depend on the data because its allowance is pathwise. In (D.12), the leading score-variance bound approaches $mB^2/4$ as $a \to 0$.

D.4 A POPULATION TANGENT CERTIFICATE FOR NUMERICAL FITTING

We specialize to the common-feature logistic construction in Appendix I. Let $w = w(N) \in \mathbb{R}^p$ satisfy $\|w\|_2 \le \ell_W$, and assume the null mean is $q_0(N) = \Lambda(w^\top \eta_0)$, where $\Lambda(u) = (1 + e^{-u})^{-1}$. Write

$$\rho(u) = \Lambda(u)\{1 - \Lambda(u)\}, \qquad L_\rho = \sup_u |\rho'(u)| = \frac{1}{6\sqrt{3}}.$$

The residual fit uses the same feature vector w. Its numerical coefficient before additive perturbation is $\widetilde{\omega}$, its final coefficient is $\bar{\omega} = \widetilde{\omega} + U_\omega$, and the prediction coefficient used in the residual fitting weights is $\bar{\eta}$.

Proposition D.18 (Implementation terms in the prediction-tangent defect)**.** *Suppose d is bounded by B and is c_d-Lipschitz. Let $H' = d$, and define*

$$F_{\omega,D}(\omega; \bar{\eta}) = \frac{1}{n}\sum_{j=1}^{n} \rho(w_j^\top \bar{\eta}) H(z_j - w_j^\top \omega) + \frac{\lambda_\omega}{2}\|\omega\|_2^2.$$

On an event of interest, suppose

$$\|\nabla F_{\omega,D}(\widetilde{\omega}; \bar{\eta})\|_2 \le r_\omega$$

and, on a coefficient set $\mathcal{T}$ containing $(\bar{\eta}, \widetilde{\omega})$,

$$\sup_{(\eta,\omega)\in\mathcal{T}} \left\| \mathbb{E}[\rho(w^\top\eta)wd(z - w^\top\omega)] - \frac{1}{n}\sum_{j=1}^{n} \rho(w_j^\top\eta)w_j d(z_j - w_j^\top\omega) \right\|_2 \le \zeta_n. \tag{D.14}$$

Population expectations here use a fresh observation with the displayed coefficients held fixed. Then the prediction-tangent defect of the final residual fit satisfies

$$\begin{aligned}\nu_n &:= \left\|\mathbb{E}[\rho(w^\top\eta_0)wd(z - w^\top\bar{\omega})]\right\|_2 \\ &\le \lambda_\omega\|\widetilde{\omega}\|_2 + r_\omega + \zeta_n + B\ell_W^2 L_\rho\|\bar{\eta} - \eta_0\|_2 + \frac{c_d\ell_W^2}{4}\|U_\omega\|_2.\end{aligned} \tag{D.15}$$

Moreover, for any prediction coefficient displacement h, the null mean of the final residual score obeys

$$\left|\mathbb{E}[\{Y - \Lambda(w^\top(\eta_0 + h))\}d(z - w^\top\bar{\omega})]\right| \le \|h\|_2\nu_n + \frac{BL_\rho\ell_W^2}{2}\|h\|_2^2. \tag{D.16}$$

Proof. Differentiation of the objective gives

$$\nabla F_{\omega,D}(\widetilde{\omega}; \bar{\eta}) = -\frac{1}{n}\sum_j \rho(w_j^\top\bar{\eta})w_j d(z_j - w_j^\top\widetilde{\omega}) + \lambda_\omega\widetilde{\omega}.$$

Thus the norm of the empirical moment is at most $\lambda_\omega\|\widetilde{\omega}\|_2 + r_\omega$. Equation (D.14) transfers this bound to the population moment at $(\bar{\eta}, \widetilde{\omega})$, at cost ζ_n.

Changing the prediction weight from $\rho(w^\top\bar{\eta})$ to $\rho(w^\top\eta_0)$ changes the moment by norm at most

$$\mathbb{E}[B\|w\|_2 L_\rho|w^\top(\bar{\eta} - \eta_0)|] \le B\ell_W^2 L_\rho\|\bar{\eta} - \eta_0\|_2.$$

Next, changing $\widetilde{\omega}$ to $\bar{\omega}$ costs at most

$$\mathbb{E}[\tfrac{1}{4}\|w\|_2 c_d|w^\top U_\omega|] \le \frac{c_d\ell_W^2}{4}\|U_\omega\|_2,$$

since $\rho \le 1/4$. The triangle inequality proves (D.15). The value of L_ρ follows by maximizing $p(1-p)|1-2p|$ over $p \in [0,1]$.

For (D.16), conditional centering replaces Y by $\Lambda(w^\top\eta_0)$. Taylor's theorem and $\Lambda'' = \rho'$ give

$$\Lambda(w^\top(\eta_0 + h)) = \Lambda(w^\top\eta_0) + \rho(w^\top\eta_0)w^\top h + R(w,h), \qquad |R(w,h)| \le \frac{L_\rho}{2}|w^\top h|^2.$$

The linear term in the score mean has absolute value at most $\|h\|_2\nu_n$. The remaining term is bounded by $(BL_\rho/2)\mathbb{E}|w^\top h|^2 \le (BL_\rho\ell_W^2/2)\|h\|_2^2$. □

Equation (D.15) separates regularization, optimization, population moment approximation, prediction coefficient error, and residual perturbation. Together with (D.16), it supplies a sufficient bound on the mean shift. In a sequence of problems with bounded B, L_ρ, ℓ_W, the conditions

$$\sqrt{m}\,\|h_m\|_2\nu_m \to_P 0, \qquad \sqrt{m}\,\|h_m\|_2^2 \to_P 0$$

make that shift negligible at the $\sqrt{m}$ scale. A first-order normal test additionally uses a conditional central limit theorem with nondegenerate limiting variance, a consistent variance estimate, and release noise negligible on this scale. The bound applies on events satisfying the displayed fitting and population moment conditions; probabilistic use takes these events to have probability tending to one. Here the coefficient error is measured relative to the population parameter η_0, while the sensitivity bounds in Appendix I measure differences between adjacent datasets.

E CALIBRATION AND STATISTICAL GUARANTEES

E.1 CALIBRATING THE COMPLETE FITTING ERROR

For calibration replication k, let the nuisance-fitting program output p_{ki}^{dp} and a_{ki}^{dp}. Its end-to-end fitting remainder is

$$R_k = \left| \sum_{i=1}^{m} (q_{ki} - p_{ki}^{\mathrm{dp}}) d(a_{ki}^{\mathrm{dp}}) \right|.$$

The synthetic label generator supplies q_{ki} during calibration. Repeating the complete fitting procedure then measures the score shift produced jointly by training resampling, regularization, nuisance randomization, and numerical fitting. These effects enter one remainder. A future audit uses its fitted predictions and the calibrated bound without accessing the generator probabilities.

Lemma E.19 (Exchangeable rank-max bound)**.** *If $R_1, \ldots, R_N, R_{K+1}$ are exchangeable extended-real random variables and $\widehat{B}_J = \max_{k \le K} R_k$, then*

$$\mathbb{P}(R_{K+1} > \widehat{B}_J) \le \frac{1}{K+1}.$$

Proof. If R_{K+1} strictly exceeds the first K values, its index is the unique location of the strict maximum among $K+1$ exchangeable variables. At most one index can be the unique strict maximum. Exchangeability makes all indices equally likely to occupy that role, so the probability for the last index is at most $1/(K+1)$. Ties only decrease this probability. □

E.2 PROOF OF THE FINITE-SAMPLE TEST

For one future run, decompose the implemented statistic as

$$\begin{aligned} T^{\mathrm{impl}} = & \sum_{i=1}^{m} (Y_i - q_i^\star) d(a_i^{\mathrm{dp}}) + \sum_{i=1}^{m} (q_i^\star - q_i) d(a_i^{\mathrm{dp}}) \\ & + \sum_{i=1}^{m} (q_i - p_i^{\mathrm{dp}}) d(a_i^{\mathrm{dp}}) + W_{\mathrm{ev}} + E_{\mathrm{sc}}. \end{aligned} \tag{E.17}$$

The first term is centered conditional on the fitted program. The third is the fitting remainder controlled by the exchangeable rank bound. The second allows a departure from the reference null probabilities, while E_{sc} accounts for numerical error. The proof combines these mean-shift allowances with a concentration bound for fresh labels and release noise.

Theorem E.20 (Proof of the finite-sample test)**.** *Assume:*

1. *the K calibration remainders and the future remainder are exchangeable under the same prespecified calibration distribution;*
2. *conditional on $\mathcal{F}_{\mathrm{pre}}$, the future labels are independent Bernoulli variables with probabilities $q_i^\star$ and $|d(a_i^{\mathrm{dp}})| \le B$;*
3. *the signed drift in Equation* (E.17) *has magnitude at most $B_{\mathrm{mis},0}$;*
4. *under the reference noise law, W_{ev} is conditionally independent of the labels and σ_{ev}-sub-Gaussian;*
5. *$|E_{\mathrm{sc}}| \le e_{\mathrm{sc}}$ deterministically;*
6. *the implemented and reference conditional noise laws, defined on a common output space, admit a coupling that disagrees with probability at most τ_{sam}. The direct case takes the implemented law itself as the sub-Gaussian reference and sets $\tau_{\mathrm{sam}} = 0$.*

Let

$$\alpha_c = \alpha_{\mathrm{tot}} - \frac{1}{K+1} - \tau_{\mathrm{sam}} > 0$$

and

$$q_A = \widehat{B}_J + B_{\mathrm{mis},0} + e_{\mathrm{sc}} + \sqrt{2\left(\frac{mB^2}{4} + \sigma_{\mathrm{ev}}^2\right)\log\frac{2}{\alpha_c}}.$$

Then $\mathbb{P}(|T^{\mathrm{impl}}| > q_A) \le \alpha_{\mathrm{tot}}$, where probability includes calibration and the future run jointly. Equality at the threshold is non-rejection.

Proof. Conditional on $\mathcal{F}_{\mathrm{pre}}$, the centered Bernoulli term for unit i lies in an interval of length $|d(a_i^{\mathrm{dp}})| \le B$. Hoeffding's lemma gives it sub-Gaussian variance proxy at most $B^2/4$. Conditional independence across labels and independence of W_{ev} imply that

$$Z = \sum_i (Y_i - q_i^\star) d(a_i^{\mathrm{dp}}) + W_{\mathrm{ev}}$$

is sub-Gaussian with proxy

$$V = \frac{mB^2}{4} + \sigma_{\mathrm{ev}}^2.$$

Hence $\mathbb{P}(|Z| > \sqrt{2V\log(2/\alpha_c)} \mid \mathcal{F}_{\mathrm{pre}}) \le \alpha_c$ under the target sampler.

On the rank event from Lemma E.19, the future fitting remainder is at most $\widehat{B}_J$. The drift and bridge assumptions bound the next two deterministic contributions by $B_{\mathrm{mis},0}$ and e_{sc}. The triangle inequality applied to Equation (E.17) then shows that rejection above q_A requires failure of the sub-Gaussian tail event, the rank event, or the sampler coupling. Their probabilities are at most α_c, $1/(K+1)$, and τ_{sam}. A union bound gives α_{tot}. Strict rejection makes equality a non-rejection. □

In the primary synthetic null, labels are generated from the declared $q_i^\star$, so the point-design drift is zero. The threshold reserves $B_{\mathrm{mis},0} = 0.30$ and the implementation-error allowance $e_{\mathrm{sc}} = 0.01$. The numerical substitution uses $\tau_{\mathrm{sam}} = 0$ for the target continuous Gaussian experiment. Section I states the additional sampler and arithmetic requirements for finite-computer validity and distinguishes them from empirical validation.

The rank argument averages over calibration and one future run under the same prespecified law. It gives the theorem's marginal guarantee. A guarantee conditional on a realized threshold, under a changed label-generating law, or across repeated deployments requires an argument for that probability statement.

E.3 Primary calibration and independent validation

E.3.1 Threshold calculation

For $K = 5{,}000$, $m = 16{,}000$, $B = 0.08$, $\sigma_{\mathrm{ev}} = 1$, $\alpha_{\mathrm{tot}} = 0.05$, and the target-law value $\tau_{\mathrm{sam}} = 0$,

$$\alpha_c = 0.05 - \frac{1}{5001} = 0.0498000399920016,$$

and

$$V = \frac{16{,}000(0.08)^2}{4} + 1 = 26.6.$$

The two-sided stochastic term is

$$Q_0 = \sqrt{2(26.6)\log\frac{2}{0.0498000399920016}} = 14.01647498391976.$$

The fresh calibration maximum is $\widehat{B}_J = 0.0942848066589036$. With $B_{\mathrm{mis},0} = 0.30$ and $e_{\mathrm{sc}} = 0.01$,

$$q_A = 0.0942848066589036 + 0.30 + 0.01 + 14.01647498391976 = 14.420759790578664.$$

Rounding upward fixes the nominal experimental rule

$$\boxed{\text{reject if and only if } |T^{\mathrm{impl}}| > 14.43.}$$

Table 1: End-to-end remainder calibration for the primary regime. Quantiles are empirical over the 5,000 retained records.

K	mean	median	$q_{.95}$	$q_{.99}$	max
5000	.01192335	.00825944	.03539623	.05426816	.09428481

Under the assumptions of Theorem E.20, this rule has marginal Type-I error at most 0.05 in the primary prespecified null regime.

Fresh validation produced $65/20{,}000 = 0.00325$ null rejections. Its one-sided 95% Clopper–Pearson upper bound is 0.00399384894393894. At the predeclared effect $\theta = 0.36$, it produced $10{,}000/10{,}000$ rejections, with one-sided lower bound 0.9997004716402234. The alternative statistic quantiles were

$$q_{.001} = 23.8492358552, \qquad q_{.01} = 27.2923831647,$$
$$q_{.05} = 30.7497564386, \qquad q_{.50} = 38.6792196012.$$

The 0.1% lower-tail margin above 14.43 was 9.4192358552. These trials measure operating performance at the selected effect. The Type-I statement comes from the rank and concentration argument, independently of the validation counts.

E.3.2 Available calibration records

The released calibration records contain all $5{,}000$ fitting remainders and reproduce their maximum and the threshold arithmetic above. Configuration files, aggregate validation outputs, and retained power statistics reproduce the reported settings and summaries.

E.4 Bounding future power before drawing evaluation labels

We can also obtain a lower bound on power from the fitted quantities alone. For each alternative transcript, a concentration bound controls the probability that fresh labels and evaluation noise erase its mean margin. An order statistic then turns these conditional bounds into a guarantee for a future transcript. Here we use the frozen C0243 empirical pool at $\theta = 0.36$ and lattice Gaussian evaluation noise with physical scale one. The nominal Monte Carlo results use their recorded NumPy-normal implementation. Appendix I separately accounts for training-record sensitivity when calibrating the complete release.

E.4.1 Conditioning before evaluation labels

To distinguish this construction from the null calibration, write U for a *pre-label transcript*. It contains the sampled feature slots, training labels, fitted nuisance functions, and their randomization, but no evaluation labels or evaluation noise. Conditional on U, the evaluation labels are independent with

$$Y_i \mid U \sim \mathrm{Bernoulli}(q_i), \qquad q_i = \Lambda(w_i^\top \eta_\circ + \theta z_i), \qquad i = 1, \dots, m.$$

Here $\eta_\circ$ is the frozen generator parameter, not a fitted prediction parameter, w_i is the fixed nuisance feature vector including its intercept, and z_i is the target-direction score. The transform slope is prespecified as $c = 1$ in this row. Define the fitted predictions and transformed residuals by

$$\widehat{p}_i = \Lambda(w_i^\top \widehat{\eta}), \qquad d_i = \mathrm{clip}_B\{c(z_i - w_i^\top \widehat{\omega})\}, \qquad C(U) = \sum_{i=1}^{m} (q_i - \widehat{p}_i) d_i.$$

All quantities on the right are determined before evaluation labels are drawn. For a centered independent evaluation noise variable Z, the ideal statistic is

$$T^\circ = C(U) + S(U) + Z, \qquad S(U) = \sum_{i=1}^{m} d_i (Y_i - q_i).$$

The deterministic implementation bridge is $|T^{\mathrm{impl}} - T^\circ| \le e_{\mathrm{sc}}$. The executed test rejects when $|T^{\mathrm{impl}}| > q_A$, where $q_A = 14.43$ in this experiment. We lower-bound power through the positive rejection tail; omitting the negative tail is conservative for this positive alternative.

E.4.2 Conditional Bernoulli and noise bounds

For $\lambda \geq 0$, conditional independence gives the exact Bernoulli log moment-generating function

$$K_U(\lambda) = \log \mathbb{E}[\exp\{-\lambda S(U)\} \mid U] \tag{E.18}$$

$$= \sum_{i=1}^{m} \log\left\{(1-q_i)e^{\lambda q_i d_i} + q_i e^{-\lambda(1-q_i)d_i}\right\}. \tag{E.19}$$

Suppose the evaluation noise satisfies $\mathbb{E}e^{tZ} \leq e^{\sigma_{\mathrm{ev}}^2 t^2/2}$ for all real t. Set $a(U) = C(U) - q_A - e_{\mathrm{sc}}$. For $a(U) > 0$, Chernoff's inequality yields

$$\mathbb{P}(T^{\mathrm{impl}} \leq q_A \mid U) \leq \inf_{\lambda \geq 0} \exp\left\{-\lambda a(U) + K_U(\lambda) + \frac{\sigma_{\mathrm{ev}}^2 \lambda^2}{2}\right\}. \tag{E.20}$$

Indeed, $T^{\mathrm{impl}} > q_A$ whenever $S(U) + Z > -a(U)$, and Markov's inequality applied to $e^{-\lambda(S+Z)}$ gives the bound. Let $H_{\mathrm{CGF}}(U)$ be one minus the right side of (E.20), and set it to zero when $a(U) \leq 0$. Then $H_{\mathrm{CGF}}(U)$ lower-bounds conditional two-sided rejection power. Any nonnegative numerical choice of λ gives a valid mathematical bound when the expression is evaluated exactly; global optimizer accuracy is not required for this implication. The supplied implementation evaluates the expression in floating point. The reported decimals are a numerical evaluation of this bound, rather than an interval-arithmetic certificate of every CGF operation.

The noise assumption applies to the exact lattice Gaussian used here. If $Z = hK$ with $\mathbb{P}(K = k) \propto \exp\{-k^2/(2s^2)\}$, then completing the square gives

$$\mathbb{E}e^{tZ} = e^{(hs)^2 t^2/2} \frac{\sum_{k\in\mathbb{Z}} e^{-(k - s^2 ht)^2/(2s^2)}}{\sum_{k\in\mathbb{Z}} e^{-k^2/(2s^2)}} \leq e^{(hs)^2 t^2/2}.$$

For completeness, the inequality follows because the periodic Gaussian sum is maximized at an integer shift: its Fourier expansion has positive coefficients proportional to $e^{-2\pi^2 s^2 \ell^2}$, and every cosine is at most one. Thus $\sigma_{\mathrm{ev}} = hs$ is a valid sub-Gaussian parameter. Equation (E.19) is exact for the Bernoulli part; the noise contribution is an upper bound on its log MGF, not an exact Gaussian-tail probability.

A simpler bound follows because the range length of $d_i(Y_i - q_i)$ is $|d_i| \leq B$. Hoeffding's lemma and independence imply the variance proxy

$$V = \frac{mB^2}{4} + \sigma_{\mathrm{ev}}^2 = \frac{16000(0.08)^2}{4} + 1 = 26.6.$$

Consequently the alternative conditional-power lower bound is

$$H_{\mathrm{H}}(U) = \begin{cases} 1 - \exp\{-a(U)^2/(2V)\}, & a(U) > 0, \\ 0, & a(U) \leq 0. \end{cases} \tag{E.21}$$

This version needs only the pre-label means and the fixed variance proxy.

E.4.3 From transcript bounds to future power

The following statement applies to either conditional bound. Its probability is over a fresh calibration sample, and its conclusion concerns the power of an independent future run from the same data-generating distribution.

Theorem E.21 (One-sided power tolerance bound)**.** *Let $U_1, \ldots, U_N$ be iid calibration transcripts and let $U_\star$ be an independent transcript with the same law. Suppose $H(U) \in [0,1]$ is a measurable lower bound on conditional rejection power for a fixed target mechanism. Fix $\gamma \in (0,1)$ and an integer $s \in \{1, \ldots, K\}$ before examining the calibration values. Write $H_{(s)}$ for their sth smallest value and define*

$$\beta_s = \sum_{j=0}^{s-1} \binom{K}{j} \gamma^j (1-\gamma)^{K-j}.$$

With calibration probability at least $1 - \beta_s$,

$$\mathbb{P}\{H(U_\star) \geq H_{(s)} \mid U_1, \ldots, U_N\} \geq 1 - \gamma, \qquad \mathrm{Power}_{\mathrm{target}} \geq (1-\gamma) H_{(s)}. \tag{E.22}$$

If the executed mechanism differs from the target by total variation at most τ uniformly conditional on the transcript, its power is at least $\max\{0, (1-\gamma)H_{(s)} - \tau\}$ on the same calibration event.

Table 2: Pre-label power certificate for the scale-one lattice evaluation configuration. Power bounds hold on a calibration event of probability at least 96.25999027% under the iid distribution in Theorem E.21.

Quantity	Value
Calibration size K / selected rank s	5000 / 17
Future excluded mass γ	0.005
Threshold q_A / arithmetic allowance e_{sc}	14.43 / 0.01
Variance proxy V	26.6
Pre-label mean order statistic $C_{(17)}$	37.87850628222726
CGF order statistic $H_{\mathrm{CGF},(17)}$	0.9999917240630811
CGF future-power lower bound	0.9949917654427657
Hoeffding order statistic $H_{\mathrm{H},(17)}$	0.999967242644362
Hoeffding future-power lower bound	0.9949674064311401
Calibration failure bound β_{17}	0.037400097282608305

Proof. Represent each calibration value as $F^{-1}(V_i)$, where F is the CDF of $H(U)$, V_i are iid uniform on $(0,1)$, and F^{-1} is the generalized inverse. This representation also covers atoms. Monotonicity of the inverse implies $H_{(s)} = F^{-1}(V_{(s)})$, and $\mathbb{P}\{H(U_\star) < F^{-1}(v)\} = F(F^{-1}(v)-) \le v$. On $V_{(s)} \le \gamma$, the future probability below $H_{(s)}$ is therefore at most γ. The probability of its complementary event is

$$\mathbb{P}(V_{(s)} > \gamma) = \mathbb{P}\{\mathrm{Bin}(K,\gamma) \le s-1\} = \beta_s.$$

Since $H \ge 0$, its future expectation is at least $(1-\gamma)H_{(s)}$ on the resulting event. Integrating the conditional-power bound proves the target-power conclusion. A total-variation bound changes the probability of the rejection event by at most τ, which proves the last claim. □

For a desired calibration failure budget δ_{cal}, choose the largest s for which $\beta_s \le \delta_{\mathrm{cal}}$. Here $K = 5000$, $\gamma = 0.005$, and $\delta_{\mathrm{cal}} = 0.05$. The resulting rank is $s = 17$:

$$\beta_{17} = 0.037400097282608305, \qquad \beta_{18} = 0.06002033490187028.$$

Thus the attained calibration confidence is at least 0.9625999027173917. The rank concerns the H values; it does not presume that the CGF bounds and the pre-label means induce the same ordering.

E.4.4 Numerical certificate and reproduction

The fresh calibration contains 5000 completed transcripts at $\theta = 0.36$, with no aborts or retries. Evaluation labels were neither generated nor read, and evaluation noise was not sampled. Four prediction-fit solver warnings were retained; the reported maximum residual was $1.4329604370246222 \times 10^{-12}$. Table 2 reports the quantities needed to reproduce the aggregation.

For target power $\pi_0 = 0.95$ and $\tau = 0$, the required order statistic is $\pi_0/(1-\gamma) = 0.9547738693467336$. For the Hoeffding version, this is equivalent to

$$C_{(17)} \ge q_A + e_{\mathrm{sc}} + \sqrt{2V \log\{1/[1-\pi_0/(1-\gamma)]\}} = 27.273996616839824.$$

Both versions exceed the target. The CGF calculation gives a future-power lower bound of 99.49917654% on a calibration event with probability at least 96.25999027%.

The supplied result archive includes all 5000 pre-label means and CGF lower bounds in `alternative_power_v4_input.json`, together with the aggregation runner and its output. Sorting the supplied bounds, taking entry 17, and multiplying by 0.995 reproduces the reported power bound. The Hoeffding calculation can be reconstructed independently from the supplied means. Our accompanying numerical check reproduces the complete aggregation output and checks the binomial rank calculation.

E.4.5 Evaluation noise and implementation

The scale-one lattice evaluation configuration uses lattice spacing $h = 2^{-40}$ and integer noise scale $s_{\mathrm{DG}} = 2^{40}$, so the physical standard-deviation parameter is $hs_{\mathrm{DG}} = 1$. Its score center is rounded to the nearest lattice point with ties to even; the rounding error is at most 2^{-41} and is included in

$e_{\mathrm{sc}} = 0.01$. The continuous replace-one sensitivity is $2B = 0.16$, and the reported integer sensitivity bound is 175921860445 lattice steps. The sampler targets the untruncated lattice Gaussian itself, so the declared target bridge uses $\tau = 0$ rather than a continuous-Gaussian approximation.

The historical C0243 evaluation mechanism used NumPy normal draws. The new power calculation uses the specified lattice evaluation mechanism, while prediction and residual-fitting randomization retain their NumPy-normal laws. Exactness refers to the evaluation sampler's mathematical random-bit target; the recorded implementation uses a domain-separated pseudorandom bit stream initialized from OS entropy. The whole-score argument in Appendix I supplies a way to account for nuisance fitting directly.

The distributional scope is given in Theorem E.21; here it uses the frozen empirical pool and generated-label distribution at $\theta = 0.36$.

F Nuisance-rate and calibration experiments

We examine how the order of the nuisance remainder determines null rejection and the threshold required for calibration. A four-point null model permits exact calculation of every score mean and variance. The same model also permits direct Monte Carlo evaluation up to $m = 10^7$ without a normal approximation to the sampled data.

F.1 Model and evaluation protocol

Let $Y \sim \mathrm{Bernoulli}(1/2)$ be independent of Z, with $\Pr(Z = 1) = 2/3$ and $\Pr(Z = -2) = 1/3$. The clipping bound is $B = 3/4$, and the population bounded-residual root is $g_H = 5/8$. For prescribed nuisance errors t, s, we compare

$$\psi^{t,s}_{\mathrm{prod}}(Y, Z) = \mathrm{clip}_{3/4}[(Y - 1/2 - t)(Z - s)], \tag{F.23}$$

$$\psi^{t,s}_{\mathrm{PBOS}}(Y, Z) = (Y - 1/2 - t)\,\mathrm{clip}_{3/4}(Z - 5/8 - s). \tag{F.24}$$

Thus both methods receive the same prediction error t and center error s, measured relative to their respective population centers. These deterministic perturbations make the nuisance-rate comparison explicit.

Each replication draws counts from a multinomial distribution with sample size m and probabilities $(1/3, 1/3, 1/6, 1/6)$ for $(Z, Y) = (1, 1), (1, 0), (-2, 1), (-2, 0)$. This produces the exact distribution of the score sum. We add independent $\xi \sim N(0, 2^2)$ and compute

$$T = \frac{\sum_{i=1}^m \psi_i + \xi}{\sqrt{m\widehat{v} + 4}}, \qquad \widehat{v} = \frac{1}{m-1}\sum_{i=1}^m (\psi_i - \overline{\psi})^2.$$

The inherited zero-center test rejects when $|T| > z_{.975}$. Within each replication, both methods use the same counts and Gaussian draw. Fixing the release-noise scale isolates the nuisance shift in this mechanism experiment; the complete private release is evaluated in the main application experiments.

For each configuration, we independently calculate $\mu = E\psi$ and $v = \mathrm{Var}(\psi)$ from the four support points. The comparison curve is

$$\Phi(-z_{.975} - \Delta) + 1 - \Phi(z_{.975} - \Delta), \qquad \Delta = \frac{m\mu}{\sqrt{mv + 4}}.$$

The branch identities $E\psi^{t,s}_{\mathrm{prod}} = -(2/3)t(1 - s)$ and $E\psi^{t,s}_{\mathrm{PBOS}} = (2/3)ts$ hold at every configuration in the reported grids. We checked them both on an 11×11 local grid with $t, s \in [-.05, .05]$ and directly at every experimental point, to absolute tolerance 10^{-12}. All reported variances and comparison curves use the exact support values.

F.2 Sample-size trajectory and matched-null calibration

The main trajectory sets $t = s = .2m^{-1/3}$ and uses 50,000 paired replications at each of nine sample sizes. Table 3 reports every sample size. PBOS rejection ranges from 4.772% to 5.148%, while

Table 3: Complete sample-size trajectory with $t = s = .2m^{-1/3}$. Rejection rates are percentages. The inherited test uses 50,000 paired replications per sample size. Matched-null rates use 30,000 independent validation draws per method after 20,000 calibration draws. Thresholds are divided by each method's own null standard deviation.

	Inherited rejection (%)		Matched-null rejection (%)		Matched-null threshold / SD	
m	Product	PBOS	Product	PBOS	Product	PBOS
1,000	10.488	4.772	5.153	5.067	2.350	1.943
3,000	13.540	4.906	4.807	5.187	2.514	1.953
10,000	17.798	4.964	4.677	4.983	2.698	1.966
30,000	23.886	5.024	4.810	5.100	2.912	1.965
100,000	32.986	5.148	4.617	4.727	3.192	1.964
300,000	44.784	5.056	4.730	4.957	3.488	1.958
1,000,000	60.722	4.864	5.193	5.013	3.854	1.955
3,000,000	76.638	5.050	4.723	5.123	4.349	1.951
10,000,000	90.814	4.858	5.103	5.083	4.932	1.946

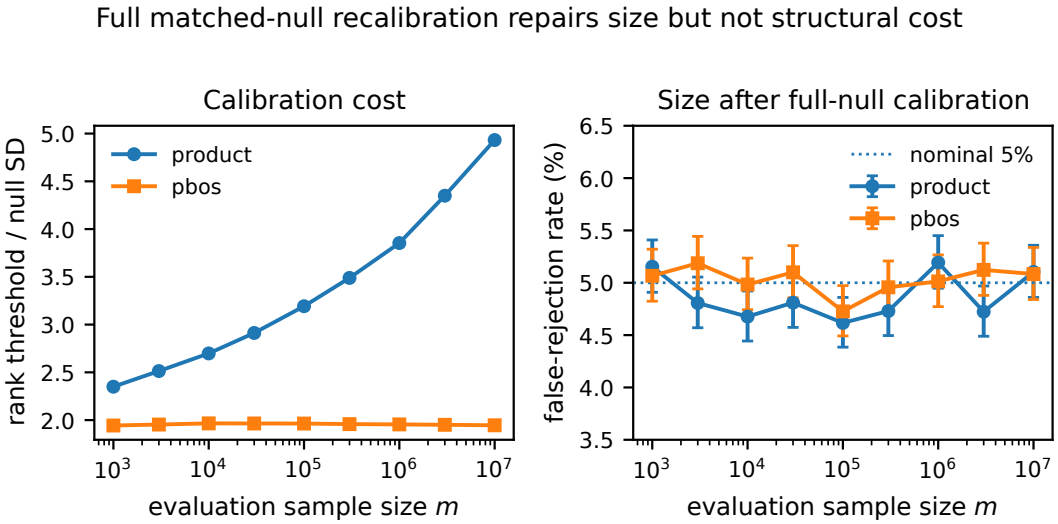


Figure 5: Matched-null critical values and independent null validation. Both methods attain empirical rejection near the nominal 5% after calibration under their full perturbed null law. Product clipping requires an increasing threshold in null-standard-deviation units; PBOS remains close to the zero-center normal critical value. Bars are two-sided Wilson 95% Monte Carlo intervals for validation conditional on the calibrated threshold.

product-clipping rejection increases from 10.488% to 90.814%. The normal comparison curve tracks the empirical trajectory. The changing rejection rates follow the corresponding standardized mean shifts: their absolute values move from .697 to 3.287 for product clipping and from .0314 to .00686 for PBOS.

We also calibrate each score under its complete, matched null distribution, including the prescribed nuisance errors and release noise. For each method and sample size, $K = 20{,}000$ independent null draws supply the absolute raw statistics $|\sum_i \psi_i + \xi|$. We take their kth order statistic, where $k = \lceil .95(K+1) \rceil = 19{,}001$, as the threshold, and assess it on 30,000 fresh null draws. The threshold in null-standard-deviation units is $q_k/\sqrt{mv+4}$. Calibration and validation use separate samples; this comparison uses the known simulation null law.

Matched-null calibration gives rejection rates near 5% for both scores. The required threshold grows from 2.350 to 4.932 null standard deviations for product clipping, while the PBOS threshold remains between 1.943 and 1.966. Figure 5 reports both threshold and size across all nine sample sizes. These are two consequences of the same mean shift: inherited calibration exposes it through null rejection, and matched-null calibration incorporates it into the critical value.

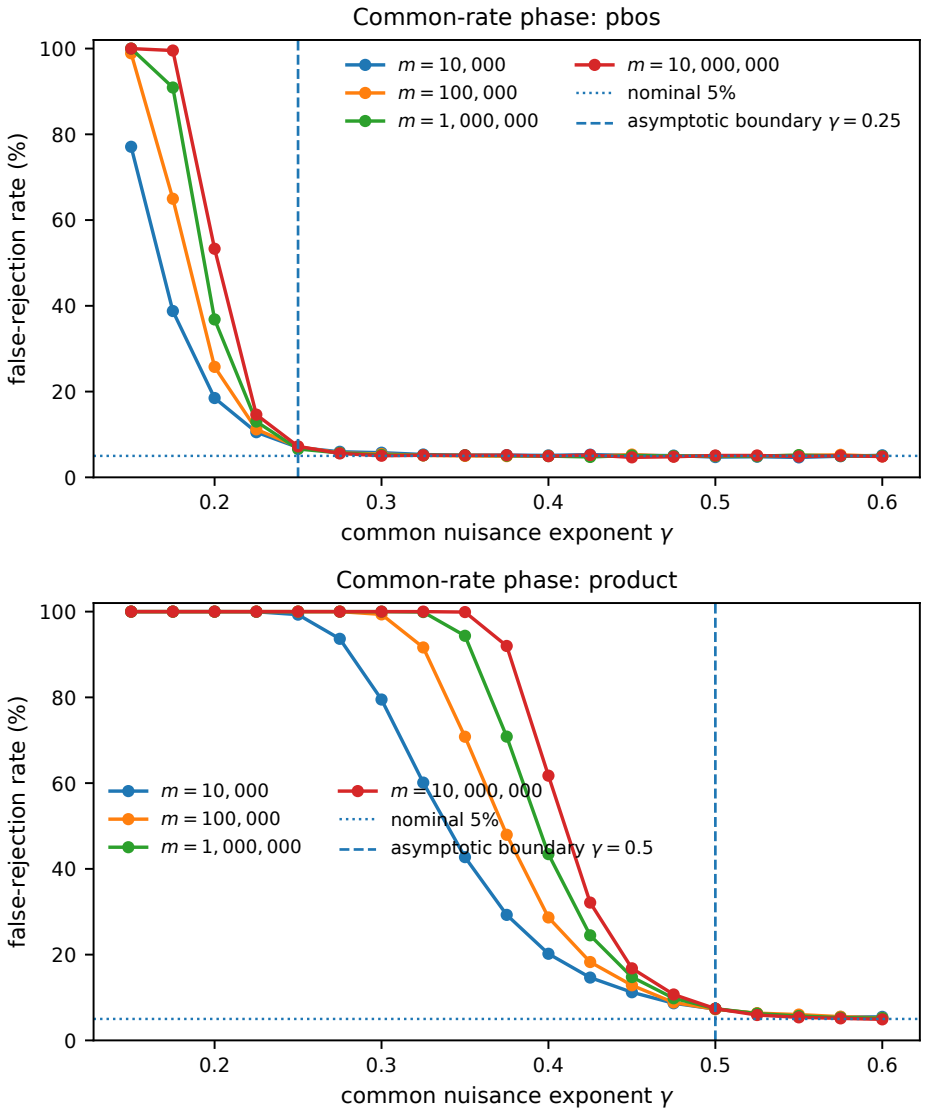


Figure 6: Common nuisance rates: $t = s = .4m^{-\gamma}$, with 20,000 paired replications per configuration. Dashed vertical lines mark the asymptotic boundaries $\gamma = 1/4$ for PBOS and $\gamma = 1/2$ for product clipping. The full rate grid includes both sides of each boundary.

F.3 Common and asymmetric nuisance rates

For the common-rate experiment, $t = s = .4m^{-\gamma}$, with $\gamma \in \{.15, .175, \ldots, .60\}$ and $m \in \{10^4, 10^5, 10^6, 10^7\}$. Each of the 76 configurations uses 20,000 paired replications. Figure 6 displays all configurations. The transition follows the second-order condition $\gamma > 1/4$ for PBOS and the first-order condition $\gamma > 1/2$ for product clipping. At $m = 10^7$, PBOS rejection is 7.225% at $\gamma = .25$ and 5.570% at $\gamma = .275$; product rejection is 7.400% at $\gamma = .50$ and 5.915% at $\gamma = .525$. At the boundary itself, the leading standardized nuisance shift remains of constant order, giving the noncentral transition seen in the curves.

The asymmetric experiment fixes $m = 10^6$ and sets $t = .4m^{-\gamma_q}$ and $s = .4m^{-\gamma_g}$. Both exponents range over $\{.10, .15, \ldots, .60\}$, yielding 121 configurations with 15,000 paired replications each. Figure 7 plots the product-minus-PBOS rejection difference in percentage points. There are 60 grid cells in the strict separation region $\gamma_q < 1/2$ and $\gamma_q + \gamma_g > 1/2$; their median difference is 92.12 percentage points and their maximum is 95.29. The full map shows how the finite-sample

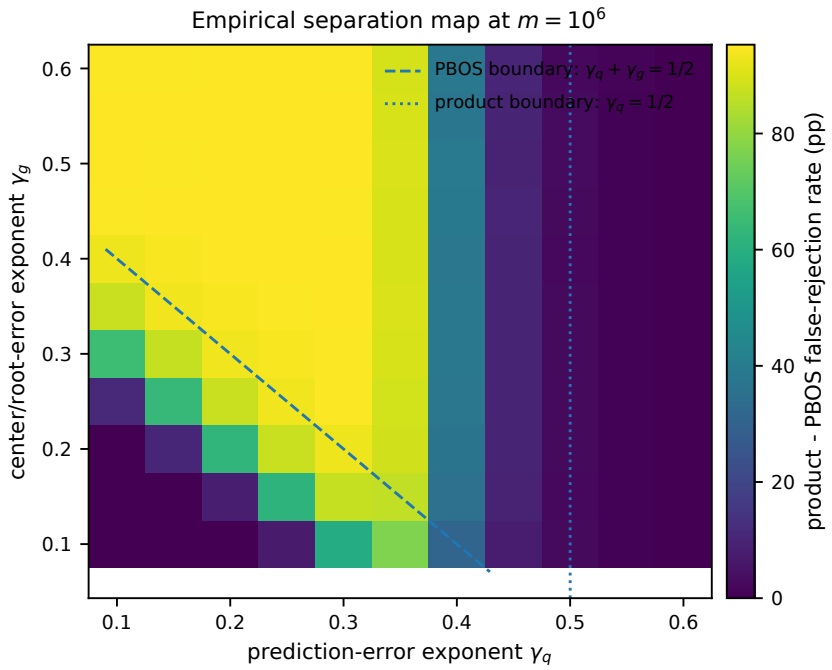


Figure 7: Asymmetric nuisance rates at $m = 10^6$. Each cell reports product-clipping rejection minus PBOS rejection, in percentage points, using 15,000 paired replications. The lines mark $\gamma_q + \gamma_g = 1/2$ and $\gamma_q = 1/2$.

transition approaches the sum-of-rates boundary for PBOS and the prediction-rate boundary for product clipping.

Reproducibility. All results in this section use the expanded experiment batch with base seed 114514: the sample-size, common-rate, asymmetric-rate, and matched-null streams use seed offsets $0, 1, 2, 10$, respectively. In total, the experiments draw 4,685,000 Monte Carlo datasets, counting each paired draw once. The supplied CSV files contain rejection counts, exact score moments, noncentralities, and Monte Carlo standard errors; sample-size and matched-null results also include two-sided Wilson 95% intervals. We reproduced all five expanded CSV files from the simulation code and verified every reported count, interval, moment, and standardized threshold. The Wilson intervals for matched-null validation condition on the realized calibration sample.

G Additional experimental results

We report the complete experiments behind the main-paper comparisons. We first describe the common data and calibration protocol, then examine score placement, effect resolution, privacy budgets, and changes to the queried direction. The final checks vary the nuisance law and corpus composition.

G.1 Data and common protocol

The source collection contains 40,000 public English texts: 20,000 items from an English Wikipedia dump (Wikimedia Foundation, 2026), 10,000 Stack Exchange items (Stack Exchange, Inc., 2014), and 10,000 abstracts from the PMC Open Access Subset (National Center for Biotechnology Information, 2026). Public pilot and main partitions contain 10,000 and 30,000 items with nonoverlapping clusters. Qwen3-Embedding-0.6B supplies the primary representation (Zhang et al., 2025). EmbeddingGemma-300M supplies the second encoder and representation geometry (Vera et al., 2025). The target identifier is `benefit_harm_v1`.

Table 4: Cross-encoder empirical power. Intervals are two-sided Wilson 95% intervals.

Model	θ	rejects	n	power	interval
Qwen3	.08	362	3000	12.07%	[10.95, 13.28]%
Qwen3	.12	1133	3000	37.77%	[36.05, 39.52]%
Qwen3	.16	2129	3000	70.97%	[69.32, 72.56]%
Qwen3	.20	2788	3000	92.93%	[91.96, 93.80]%
Qwen3	.24	2972	3000	99.07%	[98.65, 99.35]%
Qwen3	.30	2999	3000	99.97%	[99.81, 99.99]%
Qwen3	.36	3000	3000	100.00%	[99.87, 100.00]%
EmbeddingGemma	.08	284	3000	9.47%	[8.47, 10.57]%
EmbeddingGemma	.12	968	3000	32.27%	[30.62, 33.96]%
EmbeddingGemma	.16	1923	3000	64.10%	[62.37, 65.80]%
EmbeddingGemma	.20	2663	3000	88.77%	[87.59, 89.85]%
EmbeddingGemma	.24	2940	3000	98.00%	[97.43, 98.44]%
EmbeddingGemma	.30	3000	3000	100.00%	[99.87, 100.00]%
EmbeddingGemma	.36	10000	10000	100.00%	[99.96, 100.00]%

The nuisance vector W_3 contains an intercept, log-transformed character length, a certainty score, and a subjectivity score. Synthetic labels obey

$$Y \mid W_3, Z_{\text{lab}} \sim \text{Bernoulli}\left[\Lambda\{W_3^\top \eta + \theta Z_{\text{lab}}\}\right], \qquad \eta = (-0.20, 0.35, -0.30, 0.25)^\top,$$

unless a cell states a different confounding multiplier or nonlinear term. Training and evaluation slots are drawn with replacement at the fixed source weights $50/25/25$. Model fitting uses the sampled training labels; a fresh sampled evaluation set supplies the released statistic.

Unless otherwise stated, each supplementary power condition has 3,000 replications and each supplementary null condition has 5,000. Every changed model, score, confounding law, privacy schedule, or resampling law receives 5,000 fresh rank-calibration replications before validation. Two-sided Wilson 95% intervals are reported for supplementary curves and tables (Wilson, 1927). The primary validation experiment instead reports one-sided Clopper–Pearson bounds (Clopper & Pearson, 1934). Results from the two experiments are not pooled.

G.2 Results across effect sizes and encoders

The Qwen threshold is 14.43 and the Gemma threshold is 14.45. Independent Gemma null validation gives $63/20{,}000 = 0.315\%$, with Wilson interval $[0.246, 0.403]\%$. The Qwen effect curve and the confounding ladder are independent Monte Carlo batches. Their $\theta = .20$ cells therefore contain $2788/3000 = 92.93\%$ and $2813/3000 = 93.77\%$ rejections under the same nominal primary configuration. The Qwen $\theta = .36$ entry in Table 4 belongs to the supplementary curve and uses 3,000 trials. The $10{,}000/10{,}000$ primary result is a separate validation experiment. The Gemma evaluation at $\theta = .36$ contains 10,000 trials.

G.3 Results under increasing confounding

Weak, moderate, and strong confounding retain the intercept -0.20 and multiply the remaining coefficients in η by 0.5, 1.0, and 1.5. PBOS evaluates $(Y-p)\,\text{clip}_B(a)$, post hoc evaluates $\text{clip}_B\{(Y-p)a\}$, and the no-residualization reference evaluates $(Y-p)\,\text{clip}_B(z)$. The oracle reference uses public population nuisance roots. Each method-level combination is calibrated under its own executed score law.

As confounding strengthens, the post-hoc threshold rises from 31.35 to 35.15, although its branch span stays fixed at $2B$. PBOS power changes by 1.67 percentage points across the ladder, while post-hoc power falls by $445/3000 = 14.83$ points. The decomposition below shows how fitting drift contributes to this difference.

Ablations: The unresidualized score is slightly more powerful in each correctly specified cell. Its multiplier retains an additional nuisance-aligned component, and it omits the ω fitting step. With a correct prediction anchor, it is still centered under conditional independence. Its power therefore

Table 5: Method ladder at $\theta = 0.20$. Intervals are two-sided Wilson 95% intervals. PBOS and post hoc form the budget-matched comparison; the other rows are structural references.

Confounding	Method	Threshold	Null rejects / rate [interval]	Power rejects / rate [interval]
Weak	PBOS	14.44	13/5000; .26% [.15,.44]	2818/3000; 93.93% [93.02,94.73]
Weak	Post-hoc clip	31.35	2/5000; .04% [.01,.15]	2663/3000; 88.77% [87.59,89.85]
Weak	No residualization	15.23	12/5000; .24% [.14,.42]	2838/3000; 94.60% [93.73,95.35]
Weak	Oracle nuisance	14.33	26/5000; .52% [.36,.76]	2792/3000; 93.07% [92.10,93.92]
Moderate	PBOS	14.43	11/5000; .22% [.12,.39]	2813/3000; 93.77% [92.84,94.58]
Moderate	Post-hoc clip	32.40	1/5000; .02% [.00,.11]	2541/3000; 84.70% [83.37,85.94]
Moderate	No residualization	15.15	7/5000; .14% [.07,.29]	2846/3000; 94.87% [94.02,95.60]
Moderate	Oracle nuisance	14.33	13/5000; .26% [.15,.44]	2786/3000; 92.87% [91.89,93.73]
Strong	PBOS	14.43	10/5000; .20% [.11,.37]	2768/3000; 92.27% [91.26,93.17]
Strong	Post-hoc clip	35.15	0/5000; .00% [.00,.08]	2218/3000; 73.93% [72.33,75.47]
Strong	No residualization	15.27	8/5000; .16% [.08,.32]	2802/3000; 93.40% [92.45,94.23]
Strong	Oracle nuisance	14.33	12/5000; .24% [.14,.42]	2754/3000; 91.80% [90.76,92.73]

describes a different multiplier and fitting problem. The oracle rows separately show the behavior when nuisance roots are supplied exactly.

G.4 Explaining the threshold difference

The executed threshold of Theorem E.20 is a sum of four terms,

$$q_A = \underbrace{\sqrt{2\left(\frac{m\,\mathrm{span}^2}{4} + \sigma_{\mathrm{ev}}^2\right)\log\frac{2}{\alpha_c}}}_{Q_0 \text{ (conditional span)}} + \underbrace{\widehat{B}_J}_{\text{fitting remainder}} + \underbrace{B_{\mathrm{mis},0} + e_{\mathrm{sc}}}_{0.31 \text{ prespecified}},$$

so each calibrated threshold in Table 5 decomposes uniquely. Only $\widehat{B}_J$ carries information about the nuisance-fitting path; Q_0 is fixed by the conditional branch span of one summand, and the allowances are constants.

Remark G.22 (Placement sets the branch span). Section C.4 establishes label-branch spans of at most B for PBOS and $2B$ for post-hoc clipping. Their squared spans enter the variance proxy in Q_0.

With $m = 16{,}000$, $\sigma_{\mathrm{ev}} = 1$, $B = 0.08$, $N = 5{,}000$, and $\alpha_{\mathrm{tot}} = 0.05$, Remark G.22 gives $Q_0 = 14.0165$ for PBOS, the no-residualization reference, and the oracle reference, against $Q_0 = 27.6349$ for the post-hoc comparator. Table 6 reports $\widehat{B}_J$ directly from the frozen calibration finalization records and places the realized null and alternative distributions alongside it. The oracle-nuisance rows return $\widehat{B}_J = 0$, as expected when the nuisance roots are exact. The realized null deviations reproduce the predicted span ratio, $9.13/4.79 = 1.91$ against the analytic 2.

Results: Similar standardized signal can lead to different rejection rates under the calibrated rules. All four scores carry nearly the same standardized directional signal, between 4.41 and 4.74, and the normal approximation to oracle-null power exceeds 99% for all four at $\theta = 0.20$. The implemented test instead uses the finite-sample threshold, whose span and remainder terms differ across scores. For PBOS the fitting remainder stays in $[0.094, 0.112]$, about seven tenths of one percent of its threshold. For the post-hoc comparator it is 3.401 at weak confounding and 7.196 at strong, a 2.1-fold response to a perturbation that leaves the signal intact. That response is consistent with the different nuisance geometry of the final scores; the decomposition also shows the substantial contribution of the conditional branch span.

Calibration cost: Private nuisance learning adds 0.094 to the PBOS threshold against the oracle-nuisance reference, while the corresponding powers are 93.77% and 92.87% at moderate confounding. And the conservativeness of the certificate is a span-and-tail effect rather than a calibration effect: the PBOS threshold stands at $14.43/4.788 = 3.01$ realized null deviations where an asymptotic two-sided 5% rule would stand at 1.96, which implies a size of 0.258% and matches the 0.325% observed in fresh null validation.

Table 6: Threshold decomposition and standardized separation at $\theta = 0.20$. Q_0 is analytic and $\widehat{B}_J$ is the exact rank maximum stored in each frozen calibration finalization; only the execution threshold q_A is rounded upward to two decimals. Null and alternative summaries are the recorded Monte Carlo moments over $5{,}000$ and $3{,}000$ replications. The standardized separation is the alternative mean in units of the realized null deviation. Oracle-null power evaluates each score at the empirical two-sided 5% null quantile of its own statistic, under a normal approximation whose accuracy the recorded quantiles support.

Confounding	Method	q_A	Q_0	$\widehat{B}_J$	null sd	alt. mean	separation	oracle-null power
Weak	PBOS	14.44	14.02	0.112	4.888	21.846	4.47	99.35%
Weak	Post-hoc clip	31.35	27.63	3.401	9.049	42.344	4.68	99.67%
Weak	No residualization	15.23	14.02	0.900	4.861	22.905	4.71	99.74%
Weak	Oracle nuisance	14.33	14.02	0	4.933	21.756	4.41	99.18%
Moderate	PBOS	14.43	14.02	0.094	4.788	21.605	4.51	99.50%
Moderate	Post-hoc clip	32.40	27.63	4.446	9.131	41.853	4.58	99.58%
Moderate	No residualization	15.15	14.02	0.823	4.841	22.923	4.74	99.72%
Moderate	Oracle nuisance	14.33	14.02	0	4.814	21.442	4.45	99.37%
Strong	PBOS	14.43	14.02	0.102	4.756	21.169	4.45	99.27%
Strong	Post-hoc clip	35.15	27.63	7.196	8.873	40.804	4.60	99.44%
Strong	No residualization	15.27	14.02	0.941	4.839	22.360	4.62	99.65%
Strong	Oracle nuisance	14.33	14.02	0	4.751	21.091	4.44	99.29%

Table 7: Privacy and utility sweep at $\theta = .20$. Each threshold is calibrated for its own total privacy budget. Brackets contain two-sided Wilson 95% intervals.

ε	Threshold	Null rejects / rate [interval]	Power rejects / rate [interval]
1	17.98	23/5000; .46% [.31,.69]	2166/3000; 72.20% [70.57,73.77]
2	15.22	15/5000; .30% [.18,.49]	2657/3000; 88.57% [87.38,89.66]
4	14.43	11/5000; .22% [.12,.39]	2813/3000; 93.77% [92.84,94.58]
8	14.27	15/5000; .30% [.18,.49]	2795/3000; 93.17% [92.21,94.02]

G.5 Results across privacy budgets

For total budget $e \in \{1, 2, 4, 8\}$, the nuisance and evaluation budgets are $(.75e, .25e, e)$ for $(\eta, \omega, \text{evaluation})$. The nuisance mechanisms compose sequentially, while evaluation acts on a disjoint split and composes in parallel. The evaluation Gaussian sum scale is $4/e$, yielding 4, 2, 1, and .5 across the four rows; each scale is at least the analytic-Gaussian minimum for its evaluation budget. The $e = 4$ row reuses the moderate-confounding PBOS baseline, while the other budgets come from independent supplementary batches with their own calibration.

Results: The largest power gain in this schedule occurs between $\varepsilon = 1$ and $\varepsilon = 4$, and the intervals at 4 and 8 overlap. The prespecified $\varepsilon = 4$ configuration lies near the empirical knee of this four-point schedule. All four null rates are below 0.5%, showing substantial calibration slack.

G.6 Results for control queries

All control cells retain the original benefit-harm direction in the label law and replace only the query direction. The nuisance-only direction is the normalized equal sum of the certainty and subjectivity directions. Descriptor permutation uses a seeded permutation of the 24 frozen benefit-harm descriptor embeddings before subtracting the final 12 from the first 12. The random control is a seeded isotropic unit direction. The hard control is the frozen public-pilot selection from $20{,}000$ target-correlated randomized-PCA candidates under a prespecified closest-geometry criterion. Sign reversal negates the frozen target direction. Every control reuses the target query's frozen centering and scaling before clipping to $[-1, 1]$; consequently, sign reversal is a constructed query control, not a full sign-equivariance test of the training and calibration pipeline.

Results: Query specificity is visible within this fixed representation cache. The nuisance-only, permuted, and random directions remain near the null. The target-correlated hard control retains partial signal, the sign-reversed query remains high, and the specified target is strongest. These results establish specificity within the constructed control family.

Table 8: Results for control queries at $\theta = .20$. The target row is the moderate-confounding baseline. Brackets contain two-sided Wilson 95% intervals.

Direction	Threshold	Null rejects / rate [interval]	Power rejects / rate [interval]
Nuisance only	18.66	3/5000; .06% [.02,.18]	2/3000; .07% [.02,.24]
Descriptor permutation	15.12	5/5000; .10% [.04,.23]	12/3000; .40% [.23,.70]
Norm-matched random	16.58	1/5000; .02% [.00,.11]	29/3000; .97% [.67,1.38]
Geometry-matched hard control	15.64	4/5000; .08% [.03,.21]	578/3000; 19.27% [17.90,20.72]
Sign reversed	15.64	3/5000; .06% [.02,.18]	2636/3000; 87.87% [86.65,88.99]
Target	14.43	11/5000; .22% [.12,.39]	2813/3000; 93.77% [92.84,94.58]

Table 9: Alignment ladder at $\theta = .20$. Brackets contain two-sided Wilson 95% intervals.

ρ	Method	Threshold	Null rejects / rate [interval]	Power rejects / rate [interval]
0	PBOS	14.43	11/5000; .22% [.12,.39]	2813/3000; 93.77% [92.84,94.58]
0	No residualization	15.15	7/5000; .14% [.07,.29]	2846/3000; 94.87% [94.02,95.60]
.30	PBOS	14.44	12/5000; .24% [.14,.42]	2506/3000; 83.53% [82.16,84.82]
.30	No residualization	15.35	9/5000; .18% [.09,.34]	2805/3000; 93.50% [92.56,94.33]
.60	PBOS	14.45	6/5000; .12% [.06,.26]	989/3000; 32.97% [31.31,34.67]
.60	No residualization	15.56	7/5000; .14% [.07,.29]	2677/3000; 89.23% [88.07,90.29]
.80	PBOS	14.45	1/5000; .02% [.00,.11]	65/3000; 2.17% [1.70,2.75]
.80	No residualization	15.69	6/5000; .12% [.06,.26]	2598/3000; 86.60% [85.33,87.77]

G.7 Results under target–nuisance alignment

Let $U = \mathrm{clip}_{[-1,1]}\{(W_{3,\mathrm{certainty}} + W_{3,\mathrm{subjectivity}})/\sqrt{2}\}$. The aligned query and label target is

$$Z_\rho = \mathrm{clip}_{[-1,1]}\left[s_\rho\{(1-\rho)Z_0 + \rho U\}\right],$$

where s_ρ restores the public-pool root-mean-square scale. The fitted residualization center is a label-free population PBOS root on the frozen public pool. Each value of ρ and each method receives a fresh calibration.

The $\rho = 0$ rows reuse the moderate-confounding baseline. The $\rho = .60$ rows come from the second supplementary batch, while $\rho = .30$ and $.80$ come from the third. At $\rho = .60$, the separately calibrated post-hoc score has threshold 32.42, null rate $0/5000$, and power $355/3000 = 11.83\%$ with Wilson interval $[10.73, 13.04]\%$. This comparator cell supports the placement contrast at the shared alignment level. The full alignment curve was run for PBOS and no residualization.

As ρ increases, more of the target score is predictable from nuisance information and less remains in A_0. The declining PBOS power is consistent with contraction of residual energy in the scientific tangent $\mathbb{E}[\rho_0 A_0 d(A_0)]$. The unresidualized multiplier retains an aligned component and weights the directional alternative differently. With a correct prediction anchor it is still centered under conditional independence; its higher power here should not be interpreted as an equivalent nuisance-orthogonal target or as a test of marginal association.

G.8 Results under misspecification and cluster resampling

The nonlinear cell augments the true logit by

$$0.75\ \mathrm{clip}_{[-1,1]}\left[\frac{\tanh\{2W_{3,\mathrm{surface}}W_{3,\mathrm{certainty}}\} + 0.0063601121}{0.9703876921}\right],$$

while the fitted nuisance remains linear in W_3. For the cluster experiment, train and evaluation slots are sampled independently by source. Template-family clusters are resampled with replacement within each source, and records are then sampled from the resulting cluster-weighted pool to preserve the $50/25/25$ source quotas.

Results: Under the omitted nonlinear interaction, the two calibrated bounded scores have overlapping power intervals. The advantage on the confounding ladder is therefore setting dependent. Cluster resampling, by comparison, retains the primary power pattern.

Cluster resampling preserves the primary pattern when the sampling law follows corpus clusters.

Table 10: Nonlinear nuisance misspecification at $\theta = .20$. Each method uses its own calibrated threshold. Brackets contain two-sided Wilson 95% intervals.

Method	Threshold	Null rejects / rate [interval]	Power rejects / rate [interval]
PBOS	18.34	1/5000; .02% [.00,.11]	2513/3000; 83.77% [82.40,85.04]
Post-hoc clip	36.50	2/5000; .04% [.01,.15]	2526/3000; 84.20% [82.85,85.46]
No residualization	19.98	2/5000; .04% [.01,.15]	2700/3000; 90.00% [88.87,91.02]

Table 11: Source-stratified cluster resampling. Brackets contain two-sided Wilson 95% intervals.

Condition	Threshold	Rejects / rate [interval]
Null	14.47	13/5000; .26% [.15,.44]
$\theta = .20$	14.47	2775/3000; 92.50% [91.50,93.39]
$\theta = .36$	14.47	3000/3000; 100% [99.87,100]

G.9 Leave-one-source-out results

Table 12: Leave-one-source-out results. Brackets contain two-sided Wilson 95% intervals.

Excluded source	Threshold	Null	Power, $\theta = .20$	Power, $\theta = .36$
Wikipedia	14.50	13/5000; .26% [.15,.44]	2957/3000; 98.57% [98.07,98.93]	3000/3000; 100% [99.87,100]
Stack Exchange	14.50	18/5000; .36% [.23,.57]	2660/3000; 88.67% [87.48,89.75]	3000/3000; 100% [99.87,100]
PMC OA	14.52	19/5000; .38% [.24,.59]	2663/3000; 88.77% [87.59,89.85]	3000/3000; 100% [99.87,100]

The strong-effect result persists after excluding each source. The medium-effect spread of nearly ten percentage points shows that source composition still affects audit resolution.

G.10 Diagnostics for the placement comparison

We use three diagnostics to examine the placement comparison more closely. The rejection margins show the distribution underlying the power estimates. Direct perturbations isolate fitted-nuisance responses, and paired simulations measure the contribution of the worst-case concentration bound.

Figure 9 shows how often the noisy score clears its own threshold. At strong confounding, more clipped-product runs lie at or below zero. This distributional view agrees with the nominal power comparison and keeps the method-specific score scales explicit.

The perturbation calculation in Figure 10 separates the two fitted functions. Along the selected certainty direction, PBOS changes little under prediction perturbation and stays centered under residual perturbation. The clipped product responds to both. The residual result agrees with the conditional-centering calculation in Section C; the prediction result is specific to this direction and its fitted moment.

Figure 11 refines the concentration term using the two score values available for each fixed evaluation input. Both methods use the same simulated samples and noise in a new batch of 3,000 alternative runs per confounding setting. Thresholds in the plot are means of the run-specific thresholds. Under strong confounding, the observed power gap falls from 18.97 percentage points with worst-case spans to 11.37 with actual spans. These values belong to this new batch; the nominal gap is 18.33 points. The refinement shows that the original concentration bound contributes to the gap, while a difference remains after tightening both rules. This public synthetic diagnostic evaluates threshold behavior; the target privacy mechanisms retain the assumptions in Section A.

G.11 Effect resolution and privacy budget

Figure 12 collects the effect and privacy curves discussed in the main paper. The privacy sweep flattens between $\varepsilon = 4$ and 8; the recorded curve is shown without imposing monotonicity.

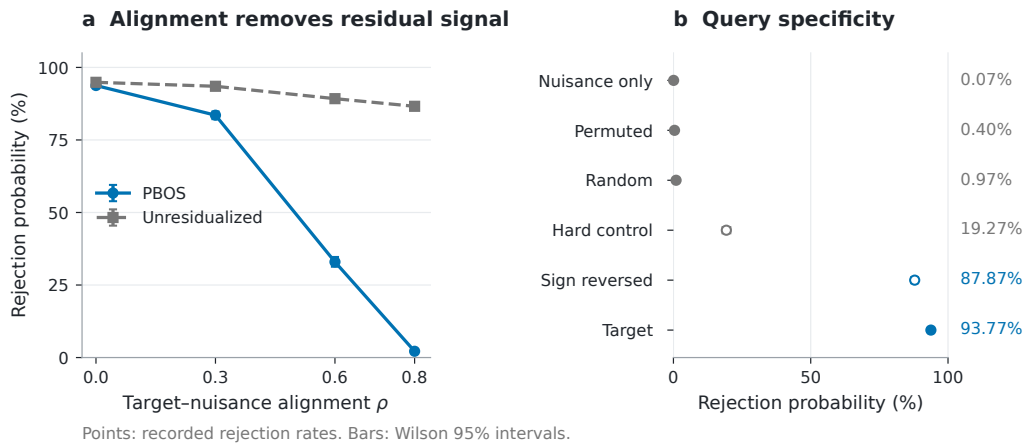


Figure 8: **Directional signal after residualization.** Left: power as the target and query become aligned with nuisance context; ρ is a construction parameter. Right: query controls under the fixed target-label model. Each point uses 3,000 runs with Wilson 95% intervals.

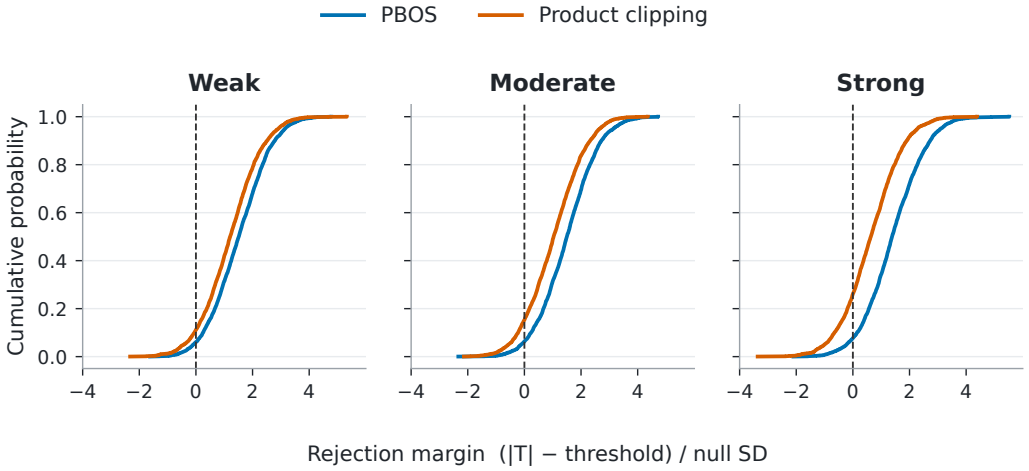


Figure 9: Nominal rejection margins under weak, moderate, and strong confounding. Each curve is the empirical CDF from 3,000 alternative runs. The horizontal coordinate is $(|T| - c)/s_0$, where c is the method's execution threshold and s_0 its empirical null standard deviation. A positive margin means rejection; the CDF at zero is the observed non-rejection fraction.

G.12 Sample size, clipping, and changes to the calibration law

Experimental design: The additional study uses synthetic binary labels on public Qwen3 English representations. It contains $13 \times 1{,}000$ calibration pairs and 78×400 validation pairs. The four displayed designs share some cells; these counts describe the underlying study rather than four independent repetitions. The comparisons use actual conditional-span diagnostic thresholds, with $K = 1{,}000$ and 400 paired validation runs per cell. This protocol differs from the primary experiment's $K = 5{,}000$ and 3,000-run supplementary power cells. Its nominal privacy schedule specifies the noise used in the comparison; the diagnostic plots do not establish a complete private release mechanism.

Effect and confounding: Figure 13 extends the placement comparison over a discrete grid of effect strengths and confounding multipliers. The largest differences occur at intermediate effects

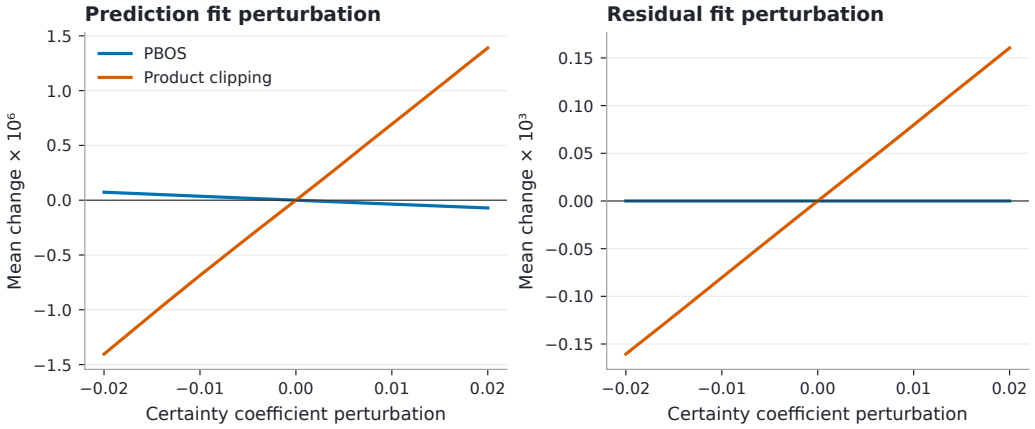


Figure 10: Direct response to a fitted certainty-coefficient perturbation under the fixed strong-confounding null label probabilities. Left: perturbing the prediction fit. Right: perturbing the residual fit. Curves show the conditional mean change relative to zero perturbation, computed without evaluation labels. The vertical scales differ: the left displays mean change multiplied by 10^6, the right by 10^3. These are deterministic calculations, without Monte Carlo intervals.

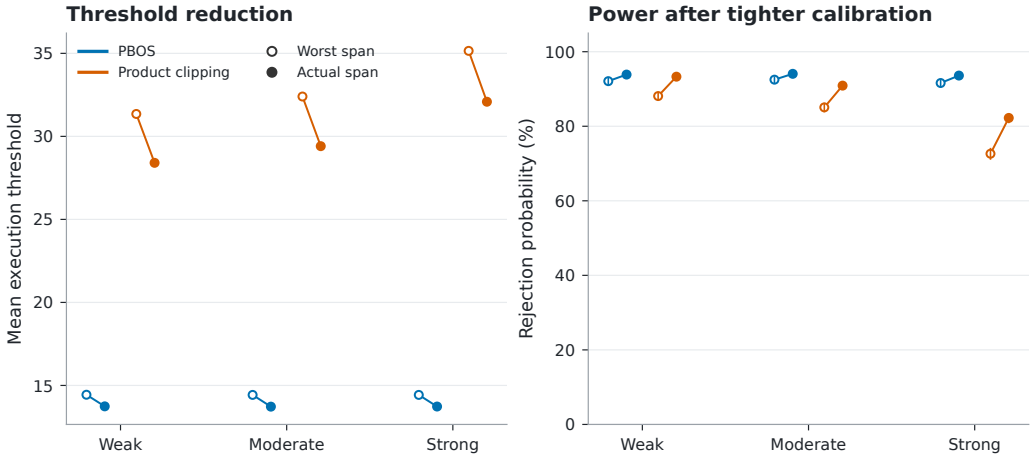


Figure 11: **Tightening both conditional-span bounds.** Empty markers use worst-case spans; filled markers use actual conditional spans. Left: mean run-specific thresholds. Right: power with Wilson 95% intervals from 3,000 paired alternative runs per setting. The methods share simulated records and noise. This batch is distinct from the nominal primary comparison.

under stronger confounding. At $\theta = 0.12$ and confounding multiplier 2, the displayed paired power difference is 21.8 percentage points. At the larger effects both methods approach saturation, so the gap narrows. The $\theta = 0$ column reports null rejection rates. Equal-width columns encode discrete settings even though the numerical effect spacings differ.

Sample efficiency: Figure 3 in the main paper varies the number of records in each split over 4,000, 8,000, 16,000, 24,000, and 32,000. Training and evaluation together use twice the displayed number of sampling slots. Under strong confounding and $\theta = 0.20$, PBOS first exceeds 90% estimated power at the 16,000 grid point; product clipping first does so at 24,000. These crossings summarize the tested sample sizes. The pointwise intervals quantify Monte Carlo uncertainty, and the connecting lines do not estimate a continuous sample-complexity curve.

Clipping strength: Figure 14 evaluates $B \in \{0.02, 0.04, 0.08, 0.12, 0.16\}$. Each setting recomputes its public residual centers, recalibrates both scores, and scales evaluation noise with B under the

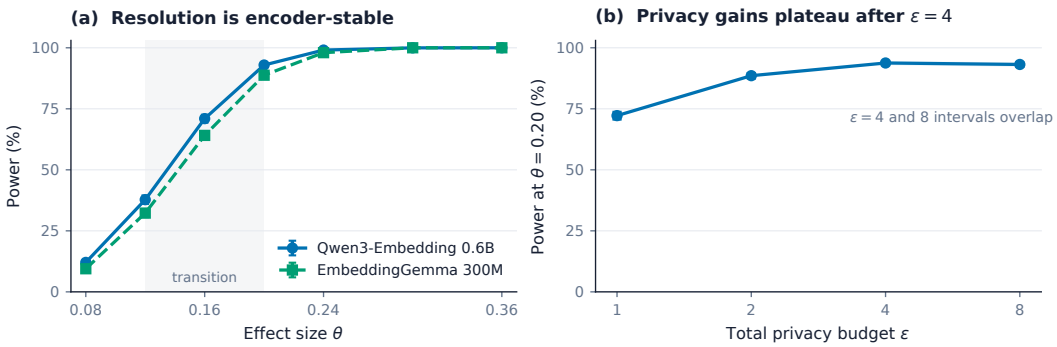


Figure 12: Power as a function of target effect and privacy budget. Left: Qwen and Gemma effect curves. Right: the PBOS privacy sweep. Points and Wilson 95% intervals correspond to the counts in the expanded experimental tables.

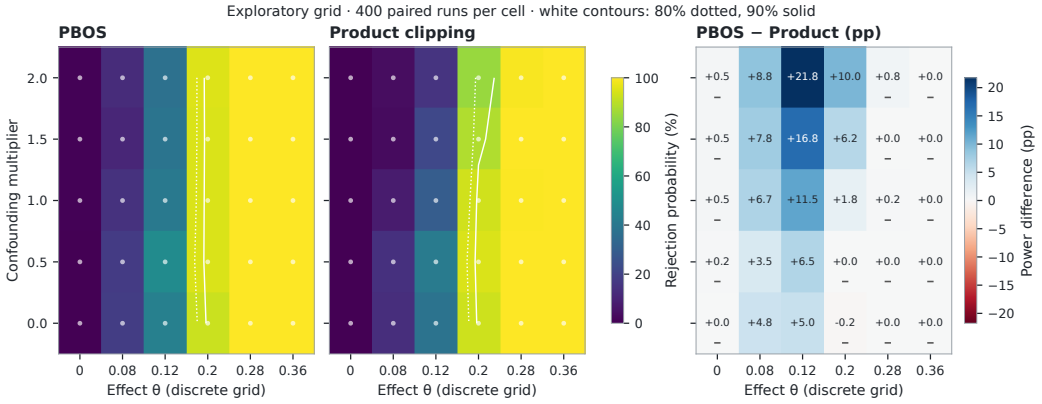


Figure 13: **Power across effect strength and confounding.** The first two panels share a rejection-rate scale; the third reports paired PBOS minus product-clipping differences in percentage points. Each cell uses 400 paired validation runs. White dotted and solid contours are grid interpolations of 80% and 90% power, respectively; dots mark observed nodes. Underlines mark cells whose pointwise approximate paired 95% interval includes zero. These intervals are not adjusted for comparisons over the grid.

nominal budget schedule. Thus, the scan compares complete choices of bound and calibration. It does not hold the score distribution and noise fixed while changing only a plotting parameter. The primary bound $B = 0.08$ remains the prespecified choice.

Across the scan, PBOS's rank maximum is small relative to its null standard deviation, while the product-clipping cost is larger. At $\theta = 0.20$, the standardized alternative means are similar, but PBOS has the lower standardized threshold. The product-clipping power declines over the larger values of B, even as its range is relaxed. This extends the main experiment's explanation: preserving signal alone is insufficient when the fitted null shift raises the threshold. The plotted mean thresholds summarize run-dependent thresholds; each rejection decision uses its own run's threshold.

Frozen calibration under drift: Figure 15 keeps training, public roots, and calibration fixed, then shifts the evaluation label law's intercept and certainty coefficient. The tested shifts are $\{-1, -0.5, 0, 0.5, 1\}$ for each coefficient. The largest observed null rejection rates on the grid are 1.00% for PBOS and 0.25% for product clipping. The blue outlines mark the unchanged-label-law reference. Each rate uses 400 runs, so a zero cell is an observed count of zero rather than a zero population rejection probability. The finite-sample theorem permits probability mismatch through its explicit D_0 bound; this empirical grid does not verify that bound over a continuous family of shifts.

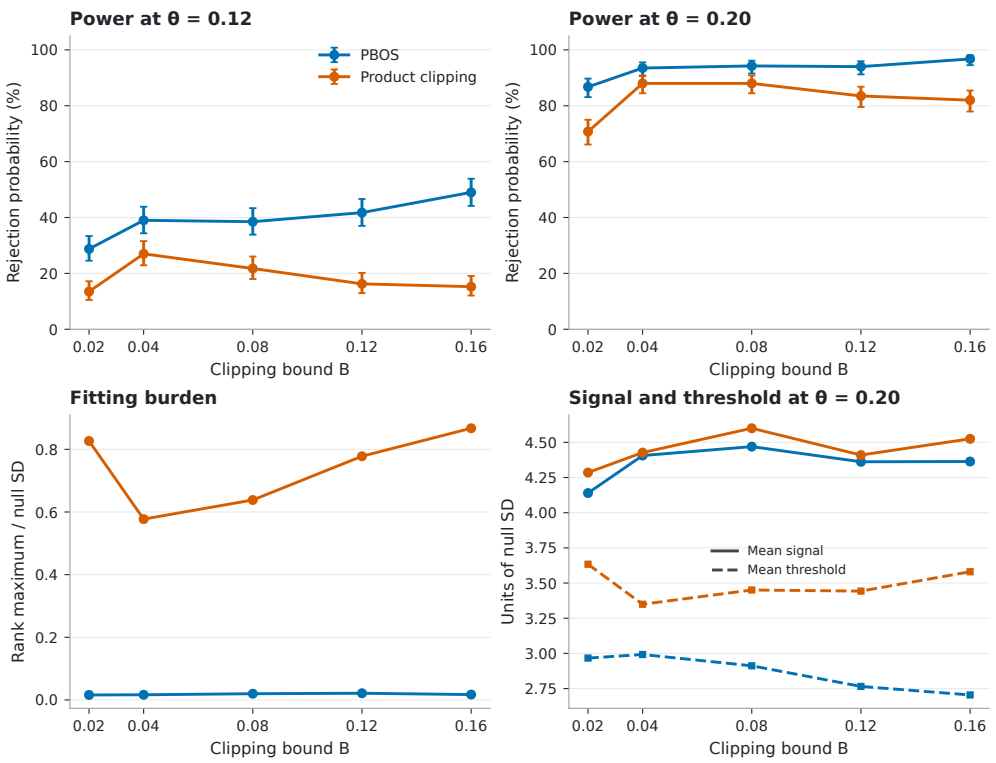


Figure 14: **Clipping strength and calibration cost.** Top: power at two effect strengths, using 400 paired runs per cell and pointwise Wilson 95% intervals. Bottom left: calibration rank maxima divided by the corresponding method-specific null standard deviation. Bottom right: alternative mean signal and mean threshold on the same standardized scale. Public centers, calibration, and evaluation-noise scales change with B under the fixed nominal budget schedule. Decisions use run-specific thresholds.

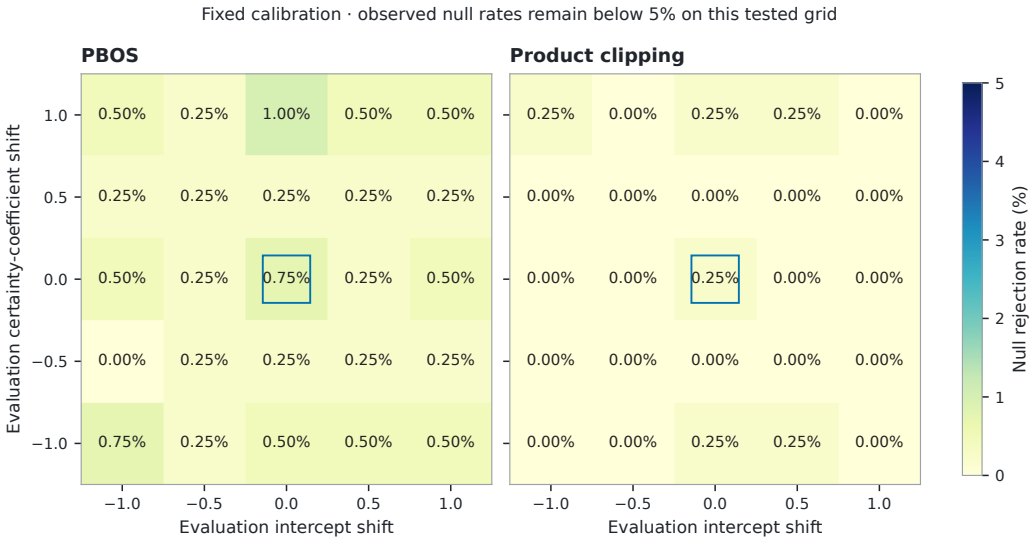


Figure 15: **Null rejection rates with calibration held fixed.** Only the evaluation label model changes, through the intercept and certainty-coefficient shifts on the two axes. Both panels use the same 0–5% color scale; blue boxes mark zero shift. Values are empirical rates from 400 paired runs per cell. Training, residual centers, and calibration remain fixed across the grid.

H EVALUATION ON NATURALLY ANNOTATED TEXT

Dataset and protocol: Civil Comments provides public comment text with crowd-annotated toxicity scores in $[0, 1]$ (Borkan et al., 2019). From its 97,320-row validation split, a deterministic label-independent ordering selects 32,000 rows and assigns 8,000 public-pilot, 12,000 training, and 12,000 evaluation comments. Toxicity is binarized at 0.5, giving positive rates of 7.90%, 8.33%, and 7.60%. The frozen benefit-harm query is paired with length, certainty, and subjectivity nuisances. Population roots are estimated only on the public-pilot split, and the primary endpoint is the held-out evaluation mean with 5,000 record-bootstrap replications.

Table 13: Held-out natural-label score summaries. Intervals are record-bootstrap 95% intervals for the mean. Overlays add the primary evaluation-noise scale to the fixed public-data score.

Method	Mean	SE	Studentized	Bootstrap interval	Sign-stable overlays
PBOS	$-.00320356$	.00018497	-17.3196	$[-.00357150, -.00284728]$	500/500
Post-hoc clip	$-.00659913$	.00039114	-16.8714	$[-.00737089, -.00582742]$	500/500
No residualization	$-.00217781$	.00018973	-11.4782	$[-.00254933, -.00181390]$	500/500

Results: The held-out PBOS interval excludes zero. Descriptively, the correlation between target and label is -0.17043, the correlation between residual and label is -0.17569, and the correlation between target and residual is 0.85779. These values show that residualization retains substantial query geometry and that the fitted held-out score exhibits association with naturally annotated labels. We interpret them as association and query-geometry diagnostics.

Interpretation: The seeded overlays describe stability to evaluation noise on these public observations. The natural-label case provides held-out evidence for a nonzero fitted-score association. Its record-bootstrap intervals do not control bias from estimating the population nuisance anchors, so they do not by themselves certify a nonzero population moment defined using q_0, g_0. A dataset-specific private release requires an independent calibration law and execution certificate.

I FINITE-COMPUTER EXECUTION

We now analyze a lattice release of PBOS that accounts for the effect of a training record on the final score. The nuisance fits remain internal. We first bound the change in their outputs under a common realization of fitting randomness, propagate that change through the evaluation sum, and add discrete-Gaussian noise to the rounded sum. This argument gives a privacy guarantee for the whole release without requiring each internal fit to be released privately.

I.1 FROM NUMERICAL FITTING TO SCORE SENSITIVITY

Fitting objectives: Let w denote the bounded nuisance vector, with $\|w\|_2 \le \ell_W$, and let n and m be the fixed training and evaluation sizes. For the construction in this section, prediction fitting minimizes the average logistic loss plus $\lambda_\eta\|\eta\|^2/2$. Residual fitting minimizes

$$F_{\omega,D}(\omega;\bar\eta) = \frac{1}{n}\sum_{j\in I_1}\rho(w_j^\top\bar\eta)H(z_j - w_j^\top\omega) + \frac{\lambda_\omega}{2}\|\omega\|^2, \qquad H'(a) = d(a),$$

where $\rho(t) = \Lambda(t)\{1-\Lambda(t)\}$ and d is nondecreasing, bounded by B, and c-Lipschitz. We use the prediction coefficient $\bar\eta$ after any internal additive randomization when forming these weights. Each fitted coefficient may receive a data-independent additive perturbation after optimization. Adjacent runs use the same perturbations for the sensitivity calculation. The fitting tolerances below refer to the objective gradient before that perturbation.

These assumptions specify a sufficient executable construction. A different residual objective or a data-dependent perturbation schedule requires its corresponding sensitivity calculation.

Lemma I.23 (Sensitivity of a numerically fitted minimizer)**.** *Let F_D be differentiable and μ-strongly convex, with minimizer $\theta^\star_D$. Suppose its numerical output satisfies $\|\nabla F_D(\widetilde\theta_D)\|_2 \le r$ on every*

admissible dataset. If adjacent exact minimizers differ by at most Δ, then

$$\|\widetilde{\theta}_D - \widetilde{\theta}_{D'}\|_2 \le \Delta + \frac{2r}{\mu}.$$

Proof. Strong monotonicity and $\nabla F_D(\theta_D^\star) = 0$ imply

$$\mu\|\widetilde{\theta}_D - \theta_D^\star\|^2 \le \langle \nabla F_D(\widetilde{\theta}_D), \widetilde{\theta}_D - \theta_D^\star\rangle \le r\|\widetilde{\theta}_D - \theta_D^\star\|.$$

Hence each numerical output is within r/μ of its minimizer. Apply the triangle inequality across the two datasets. □

Proposition I.24 (Sensitivity of the complete PBOS sum)**.** *Use the objectives above, fixed split membership, replace-one adjacency, and gradient-residual tolerances r_η, r_ω. Conditional on any fixed internal perturbations, define*

$$\Delta_\eta = \frac{2\ell_W}{n\lambda_\eta} + \frac{2r_\eta}{\lambda_\eta}, \tag{I.25}$$

$$\Delta_\omega = \frac{1}{\lambda_\omega}\left(\frac{B\ell_W}{2n} + B\ell_W^2 L_\rho \Delta_\eta\right) + \frac{2r_\omega}{\lambda_\omega}, \qquad L_\rho = \frac{1}{6\sqrt{3}}. \tag{I.26}$$

For

$$S(D) = \sum_{i\in I_2}\{Y_i - \Lambda(w_i^\top \bar{\eta})\}d(z_i - w_i^\top \bar{\omega}),$$

a training replacement changes S by at most

$$\Delta_{\mathrm{tr}} = m\left(\frac{B\ell_W}{4}\Delta_\eta + c\ell_W\Delta_\omega\right).$$

An evaluation replacement changes it by at most $2B$. Thus the whole-score sensitivity is bounded by $\Delta_{\mathrm{final}} = \max\{\Delta_{\mathrm{tr}}, 2B\}$.

Proof. One logistic-loss gradient has norm at most ℓ_W. Replacing one training record changes the average gradient by at most $2\ell_W/n$, and strong convexity gives the exact-minimizer sensitivity in (I.25). Lemma I.23 accounts for the two numerical residuals. Adding the same perturbation to both coefficients leaves their difference unchanged.

For a fixed prediction coefficient, one weighted residual-loss gradient has norm at most $B\ell_W/4$. Its replacement changes the average gradient by at most $B\ell_W/(2n)$. Changing the prediction coefficient changes a weight by at most $L_\rho\ell_W\Delta_\eta$, and hence changes the average residual gradient by at most $B\ell_W^2 L_\rho\Delta_\eta$. Dividing the sum of these bounds by λ_ω and allowing both optimization residuals yields (I.26). The bound L_ρ follows by maximizing $p(1-p)|1-2p|$ on $[0,1]$.

For any fixed evaluation record, the logistic derivative bound and the Lipschitz property of d give

$$|p - p'| \le \frac{\ell_W}{4}\Delta_\eta, \qquad |d - d'| \le c\ell_W\Delta_\omega.$$

Add and subtract $(Y - p')d$ in the difference of contributions. Since $|d| \le B$ and $|Y - p'| \le 1$, its magnitude is at most $B\ell_W\Delta_\eta/4 + c\ell_W\Delta_\omega$. Summation proves the training bound. The evaluation bound follows from the contribution range $[-B, B]$. An adjacent record belongs to only one of the fixed splits. □

Lemma I.25 (Integrating internal randomness)**.** *Let U have a data-independent distribution. If the conditional release $M_u(D)$ is (ε,δ)-DP for every fixed u, with the same parameters, then $M_U(D)$ is (ε,δ)-DP.*

Proof. For every measurable event A, integrate $\mathbb{P}\{M_u(D) \in A\} \le e^\varepsilon \mathbb{P}\{M_u(D') \in A\} + \delta$ with respect to the common law of U. □

I.2 A DISCRETE-GAUSSIAN RELEASE

Choose a public lattice spacing $h > 0$, round once with a fixed nearest-integer rule, and release

$$Q_h(D) = \mathrm{round}_{\mathrm{even}}\{\widetilde{S}(D)/h\}, \qquad M(D) = h\{Q_h(D) + \xi\}, \qquad \mathbb{P}(\xi = k) \propto e^{-k^2/(2s^2)}. \tag{I.27}$$

The integer-valued noise is drawn independently using an exact discrete-Gaussian sampler (Canonne et al., 2020). The public decision is a fixed-threshold post-processing of M. Computation and serialization retain the lattice value exactly.

Proposition I.26 (Privacy of the lattice release)**.** *Under Proposition I.24 and the uniform score-error bound, a valid integer sensitivity is*

$$D_{\mathbb{Z}} = \left\lceil \frac{\Delta_{\mathrm{final}} + 2e_{\mathrm{num}}}{h} \right\rceil + 1.$$

Suppose the discrete-Gaussian privacy profile at sensitivity $D_{\mathbb{Z}}$, scale s, and privacy parameter ε is at most δ. Then (I.27)*, together with its fixed-threshold decision, is (ε, δ)-DP for the complete training and evaluation dataset.*

Proof. The two score approximations contribute at most $2e_{\mathrm{num}}$ to adjacent differences. Nearest-lattice rounding contributes at most one further lattice step to the difference of their integer centers. Conditional on the internal fitting randomness, the discrete-Gaussian mechanism therefore satisfies the stated sensitivity-based privacy bound (Canonne et al., 2020, Theorem 7). Lemma I.25 integrates the fitting randomness. Scaling, exact serialization, and the fixed decision are data-independent post-processing. □

Numerical substitution: Set $n = m = 16000$, $\ell_W = 2$, $B = 0.08$, $c = 1$, $(\lambda_\eta, \lambda_\omega) = (0.4, 0.2)$, and $r_\eta = r_\omega = 10^{-8}$. The formulas give

$$\Delta_\eta = 0.00062505, \qquad \Delta_\omega \le 0.000121332742869, \qquad \Delta_{\mathrm{final}} \le 4.28267977180.$$

For $e_{\mathrm{num}} \le 10^{-8}$, we may conservatively use 4.282711 as the bound on $\Delta_{\mathrm{final}} + 2e_{\mathrm{num}}$. Taking $h = 2^{-40}$ and $s = \lceil 5.12/h \rceil$ gives

$$D_{\mathbb{Z}} = 4708890542906, \qquad s = 5629499534214.$$

For clarity, the integer-scale privacy calculation is as follows. Write $\bar{\Phi} = 1 - \Phi$ and

$$t_\pm = \frac{\varepsilon s^2}{D_{\mathbb{Z}}} \pm \frac{D_{\mathbb{Z}}}{2}.$$

At $\varepsilon = 4$, these thresholds are positive and nonintegral, so the discrete-Gaussian upper bound in Canonne et al. (2020, Equation (5)) gives

$$\delta \le \bar{\Phi}\left(\frac{\lfloor t_- \rfloor}{s}\right) - \left(1 - \frac{1}{\sqrt{2\pi}s + 1}\right) e^4 \bar{\Phi}\left(\frac{\lfloor t_+ \rfloor}{s}\right) \approx 9.585394 \times 10^{-7} < 10^{-6}. \tag{I.28}$$

Thus the specified residual and arithmetic conditions are sufficient for a $(4, 10^{-6})$ release. These are parameter substitutions into the construction, not privacy claims retroactively assigned to the NumPy-normal runs. Exact sampling here refers to the mathematical random-bit sampler; a pseudorandom implementation has the corresponding computational randomness assumption.

I.3 TRANSFERRING CALIBRATION THROUGH FITTING-ERROR BOUNDS

The same statistical theorem applies to this release. The discrete Gaussian is sub-Gaussian with physical parameter hs, as shown in Appendix E.4. We can therefore take the implemented noise law itself as the reference, with $\tau_{\mathrm{sam}} = 0$. Rounding and score approximation contribute $e_{\mathrm{sc}} = e_{\mathrm{num}} + h/2$.

Lemma I.27 (Calibration from reference envelopes)**.** *Let $R_1^\star, \ldots, R_N^\star, R_{K+1}^\star$ be exchangeable remainders from a common reference fitting program. Suppose available calibration values and deterministic bounds satisfy*

$$R_k^\star \le R_k + e_k \quad (k \le K), \qquad R_{K+1}^{\mathrm{exec}} \le R_{K+1}^\star + e_f.$$

Then, for $B_b^{\mathrm{exec}} = \max_{k \le K}(R_k + e_k) + e_f$,

$$\mathbb{P}\{R_{K+1}^{\mathrm{exec}} > B_b^{\mathrm{exec}}\} \le \frac{1}{K+1}.$$

Proof. On the rank event $R^\star_{K+1} \le \max_{k\le K} R^\star_k$, the assumed inequalities imply $R^{\mathrm{exec}}_{K+1} \le B^{\mathrm{exec}}_b$. Lemma E.19 bounds the complement of that event. □

For the two objectives above, one useful reference uses their exact minimizers and the same internal additive perturbations. With common training data and randomness, define

$$a_\eta = \frac{r_\eta}{\lambda_\eta}, \qquad a_\omega = \frac{r_\omega + B\ell_W^2 L_\rho a_\eta}{\lambda_\omega}.$$

The proof of Proposition I.24, now comparing a numerical fit with its exact reference on the same dataset, gives

$$E_{\mathrm{fit}}(r_\eta, r_\omega) = m\left(\frac{B\ell_W}{4}a_\eta + c\ell_W a_\omega\right). \tag{I.29}$$

The same bound holds for the absolute remainder difference because $|q_i - p_i| \le 1$. Hence e_k can combine this expression at the recorded calibration residuals with a bound on the computation of R_k; e_f uses the future uniform residual contract. Reference exchangeability requires the same feature, label, and internal-randomization laws in every run.

Corollary I.28 (Statistical threshold for the lattice release)**.** *Under Theorem E.20 and Lemma I.27, the lattice release uses the threshold*

$$q^{\mathrm{exec}}_A = B^{\mathrm{exec}}_b + B_{\mathrm{mis},0} + e_{\mathrm{num}} + h/2 + \sqrt{2\left(\frac{mB^2}{4} + (hs)^2\right)\log\frac{2}{\alpha_{\mathrm{tot}} - 1/(K+1)}}.$$

Strict rejection above this threshold has marginal Type-I error at most α_{tot}.

Proof. Use the lattice noise as its own sub-Gaussian reference, the envelope lemma for the rank failure, and $|\widetilde{S} - S| + h/2$ for arithmetic error in Theorem E.20. □

I.4 Transferring the power calculation

The Hoeffding power bound gives a way to reuse pre-label means once their fitting-error envelopes are available. Let C_k be the retained alternative means, $C^\star_k$ their exchangeable exact-reference counterparts, and $|C_k - C^\star_k| \le e_k$. Suppose every future executed mean is at least its reference mean minus e_f. For a rank $\kappa \in \{1, \dots, K\}$, set $L_k = C_k - e_k$ and let $L_{(\kappa)}$ denote their κth order statistic. Define $\beta_\kappa = P\{\mathrm{Bin}(K, \gamma) \le \kappa - 1\}$. With

$$V_{\mathrm{exec}} = mB^2/4 + (hs)^2, \qquad a = L_{(\kappa)} - e_f - q^{\mathrm{exec}}_A - e_{\mathrm{num}} - h/2,$$

the argument of Theorem E.21 gives, on a calibration event with probability at least $1 - \beta_\kappa$,

$$\mathrm{Power}_{\mathrm{exec}} \ge (1-\gamma)\left[1 - \exp\left\{-\frac{(\max\{a, 0\})^2}{2V_{\mathrm{exec}}}\right\}\right]. \tag{I.30}$$

J Experimental reproduction

The accompanying source package includes mechanism configurations, calibration records, and validation summaries. The supplementary simulation batches contain 523,000 completed Monte Carlo replications and no aborted cell: 170,000 calibration records and 353,000 validation records, split as 273,000, 120,000, and 130,000 across the three batches. Reproduction of the primary result begins by recomputing the 5,000 fitting remainders, their maximum, and the threshold in Theorem E.20. The natural-label package uses SHA-256 orderings for label-independent selection and split assignment and records 32,000 public comments plus 1,500 seeded overlays across the three scores.

The primary experimental configuration is

$$n_{\mathrm{train}} = n_{\mathrm{eval}} = 16{,}000, \quad B = 0.08, \quad (\lambda_\eta, \lambda_\omega) = (0.4, 0.2), \quad (\varepsilon_\eta, \varepsilon_\omega, \varepsilon_{\mathrm{ev}}) = (3, 1, 4).$$

The experiment uses random seed 114514; separate derived seeds govern calibration, null validation, and power validation. The rejection threshold was fixed from independent end-to-end remainder calibration before either validation sample was generated.