"Calibeating" Prediction-Powered Inference

Lars van der Laan
Department of Statistics
University of Washington

Mark van der Laan*
Department of Statistics and
Division of Biostatistics
Center for Targeted Machine Learning
and Causal Inference
University of California, Berkeley

Abstract

Prediction-powered inference (PPI) uses black-box predictions to improve estimation when labeled data are scarce but unlabeled data are abundant. However, prediction scores may be miscalibrated and need not provide efficient regression adjustments. We study semisupervised mean estimation and propose *calibrated prediction-powered inference*: calibrate the prediction scores using the labeled sample, then estimate the mean by averaging the calibrated predictions over the pooled covariate sample. This procedure requires no retraining, has a simple plug-in form, and admits an exact augmented inverse-probability weighting (AIPW) representation. We show that linear calibration is first-order equivalent to PPI++. For isotonic calibration, we establish asymptotic normality, valid Wald inference, and "*calibeating*" guarantees: isotonic calibration improves both prediction accuracy and the efficiency of the resulting regression adjustment within the class of monotone transformations. Moreover, further post-processing of the calibrated score cannot improve first-order efficiency. We also show that the original PPI estimator is a special case of AIPW and can be inefficient even when the predictions are accurate. Simulations, benchmark reproductions, and an LLM-evaluation application show that calibrated estimators often outperform PPI and are competitive with AIPW and PPI++.

1 Introduction

We consider semisupervised mean estimation with one labeled dataset $\{(X_i, Y_i)\}_{i=1}^n$ and one unlabeled dataset $\{\widetilde{X}_j\}_{j=1}^N$, where outcomes are observed only for a subset sampled completely at random. This is a classical missing-outcome problem: covariates are available for many individuals, but the outcome of interest is observed only for a smaller subset. Such settings arise when gold-standard outcomes are expensive to collect but prediction scores are readily available, for example in randomized trials, biobank analyses, and model-assisted scientific pipelines [89, 43, 115, 1, 118].

The goal is to use the larger unlabeled sample to improve estimation of a population mean while accounting for prediction error. This problem has long been studied in missing-data and semiparametric statistics, where AIPW, TMLE, and related debiased estimators combine flexible prediction with valid inference [75, 103, 105, 104, 18]. Closely related ideas have recently appeared in the machine-learning literature as prediction-powered inference [1, 2, 118]. Similar ideas also arise in covariate adjustment for randomized trials, where machine-learning predictors trained on auxiliary data can improve precision in small samples [35, 83, 61, 58, 89, 4, 42].

A standard approach is to debias a plug-in estimator based on a prediction score $m(X)$. The AIPW estimator averages $m(X)$ over the pooled covariate sample and adds the average labeled residual

*Targeted ML Solutions, targetedml.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

$Y - m(X)$ [75]. The original PPI estimator instead averages $m(X)$ only over the unlabeled sample [1] and is generally statistically inefficient. Both estimators remain valid when m is misspecified, but their efficiency depends on how well $m(X)$ serves as a regression adjustment. A score may rank observations well while being poorly calibrated, so its numerical values can remain misaligned with the conditional mean of Y.

This gap motivates post-hoc calibration. Rather than correcting only the marginal bias $\mathbb{E}\{Y - m(X)\}$, we use the labeled sample to transform $m(X)$ so that its values better match the outcome scale [113, 114, 65]. Ideally, the transformed score $q(X)$ satisfies $q(X) \approx \mathbb{E}\{Y \mid q(X)\}$: among observations with the same calibrated value, the average outcome is close to that value. We calibrate the black-box score on the labeled sample and average the calibrated predictions over the pooled covariate sample. We call this procedure *calibrated prediction-powered inference*. Calibration can improve both predictive accuracy and the efficiency of the resulting mean estimator.

Contributions. Our contributions are threefold. First, we develop a calibrated-PPI framework that includes standard post-hoc rules such as linear [60, 53], histogram binning [113, 30], and isotonic calibration [114, 65]. These procedures can be added to an existing prediction pipeline without retraining the original score. Second, we show that the resulting plug-in estimator is exactly an AIPW estimator based on the calibrated score. This identity explains why calibration can improve efficiency without changing validity: calibration changes the regression adjustment, while the estimator retains its AIPW form. Third, we establish asymptotic theory for the main calibrated estimators. Linearly calibrated PPI is first-order equivalent to unconstrained PPI++, while isotonic calibration yields asymptotic linearity and valid Wald inference. Isotonic calibration achieves the smallest asymptotic variance among nondecreasing transformations of the original score. Further post-processing of the fitted isotonic score yields no additional first-order variance reduction. Following the terminology of Foster and Hart [25], we call improvement through calibration *calibeating*. Here, these improvements concern squared prediction error and the first-order variance of semisupervised mean estimation.

1.1 Related work

Several recent PPI estimators can be viewed as AIPW estimators with different choices of regression-adjustment score. Cross-PPI uses cross-fitting to avoid overfitting bias, while PPI++ chooses an empirical scaling of the score to improve efficiency. This view connects PPI to classical missing-data, semiparametric, debiased machine-learning, and covariate-adjustment methods for mean estimation [75, 103, 105, 104, 18, 89, 42]. Semiparametric efficiency theory shows that precision depends on the quality of the regression-adjustment score, while unbiasedness does not require the score to be correctly specified [75, 5, 74, 112].

The closest precursor to our approach is calibrated debiased machine learning, which calibrates outcome-regression and debiasing nuisance estimators inside AIPW procedures [100]. That work targets robustness to nuisance misspecification in nonrandomized settings; here, random sampling already gives unbiasedness for any fixed score, so calibration is used to improve efficiency and justify post-processing a fixed score. Recalibrated prediction-powered inference (RePPI) modifies the estimating equation with a flexible machine-learned correction to target full semiparametric efficiency [45]. In contrast, we post-process the fixed prediction score using standard calibration maps and characterize the resulting guarantees for semisupervised mean estimation. Appendix B gives the full related-work discussion.

2 Two-Sample Semisupervised Setup

2.1 Data structure and notation.

We consider the standard prediction-powered inference setting with a small labeled sample and a larger unlabeled sample:

$$\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^{n}, \qquad \mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^{N}.$$

The labeled observations are i.i.d. from P_0 on $\mathcal{X} \times \mathbb{R}$, while the unlabeled observations are i.i.d. from the X-marginal $P_{0,X}$ of P_0. The estimand is the population mean

$$\psi_0 := \mathbb{E}_{P_0}[Y], \qquad \mu_0(x) := \mathbb{E}_{P_0}[Y \mid X = x].$$

Let $m : \mathcal{X} \to \mathbb{R}$ denote a black-box prediction model, trained externally or on an independent split, and write $T = m(X)$ for its scalar prediction.

For functions $a : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}$, define the empirical means over the labeled and unlabeled samples by

$$\mathbb{P}_n^L a := \frac{1}{n}\sum_{i=1}^{n} a(X_i, Y_i), \qquad \mathbb{P}_N^U b := \frac{1}{N}\sum_{j=1}^{N} b(\widetilde{X}_j).$$

Let $M := n + N$ and $\rho_n := n/M$. For $q : \mathcal{X} \to \mathbb{R}$, define the pooled covariate empirical mean

$$\mathbb{P}_M^X q := \frac{1}{M}\Big\{\sum_{i=1}^{n} q(X_i) + \sum_{j=1}^{N} q(\widetilde{X}_j)\Big\} = \rho_n \mathbb{P}_n^L\{q(X)\} + (1-\rho_n)\mathbb{P}_N^U\{q(\widetilde{X})\}.$$

Assumption 1 (Two-sample design)**.** The following hold:

(i) $\mathcal{D}_L$ is i.i.d. from P_0, $\mathcal{D}_U$ is i.i.d. from $P_{0,X}$, and the two samples are independent.

(ii) Y has finite second moment under P_0.

(iii) As $n \to \infty$, the labeled fraction $\rho_n = n/(n+N)$ satisfies $\rho_n \to \rho_0$ for some $\rho_0 \in [0, 1)$.

2.2 Review of AIPW, PPI, and semiparametric efficiency

We first review how AIPW, PPI, and PPI++ use black-box predictions to improve mean estimation, and why calibrating these predictions can improve efficiency.

The AIPW class. For any regression-adjustment score $q : \mathcal{X} \to \mathbb{R}$ with finite second moment, define

$$\widehat{\psi}(q) := \mathbb{P}_M^X q + \mathbb{P}_n^L\{Y - q(X)\}. \tag{1}$$

This is the AIPW estimator in the unrestricted two-sample model [75]. For any fixed q, the estimator is unbiased, regardless of whether the score is calibrated. The same holds conditionally for a score trained on independent data. Thus the choice of regression adjustment affects efficiency rather than unbiasedness.

Asymptotic variance and efficiency. For a fixed score q, the asymptotic variance satisfies (see Appendix C)

$$\mathrm{AVar}\Big\{\sqrt{n}\,(\widehat{\psi}(q) - \psi_0)\Big\} = V_{\text{eff}} + (1-\rho_0)\mathbb{E}_{P_0}\left[(q^\circ(X) - \mu_0(X))^2\right],$$

where $q^\circ(X) := q(X) - \mathbb{E}_{P_{0,X}}\{q(X)\} + \psi_0$ is the recentered score and

$$V_{\text{eff}} = \rho_0 \mathrm{Var}\{\mu_0(X)\} + \mathbb{E}[\mathrm{Var}(Y \mid X)]$$

is the semiparametric efficiency bound [8]. Thus the efficiency loss is proportional to the squared prediction error of the recentered score. The efficiency bound is attained when q equals the outcome regression μ_0 up to an additive constant.

AIPW, PPI, and empirical-efficiency maximization. AIPW using the raw prediction score takes $q = m$. The original PPI estimator of Angelopoulos et al. [1] is

$$\widehat{\psi}_{\text{PPI}} = \mathbb{P}_N^U m + \mathbb{P}_n^L\{Y - m(X)\},$$

which corresponds to AIPW with score $q = m/(1-\rho_n)$. Thus AIPW and PPI differ only in how they scale the original score. In particular, PPI can be less efficient than AIPW when m already provides an accurate regression adjustment.

Rather than fixing the scaling coefficient, PPI++ [2] chooses it to minimize an empirical estimate of the asymptotic variance. This is an instance of empirical efficiency maximization [83]. For unrestricted scaling, the population-optimal coefficient is

$$\lambda^\star = \frac{\mathrm{Cov}\{Y, m(X)\}}{\mathrm{Var}\{m(X)\}}.$$

PPI++ therefore adapts the scaling of the score, but cannot correct general nonlinear miscalibration. More generally, empirical efficiency maximization can be applied to richer classes of regression adjustments [83, 103]; see Appendix D. This motivates our approach: use post-hoc calibration to improve the original score, while retaining the same AIPW estimator.

3 Calibrated Prediction-Powered Inference

The previous section shows why calibration matters: AIPW is unbiased for any fixed regression adjustment, but its efficiency depends on how well that adjustment predicts the outcome. In particular, using the raw prediction score $m(X)$ can be inefficient. In applications, this score may rank individuals well even when its numerical values are poorly aligned with the outcome.

Calibration addresses this problem by transforming $m(X)$ into a new score $q(X)$ whose values better reflect the outcome. Ideally, a calibrated score satisfies

$$q(X) \approx \mathbb{E}_{P_0}\{Y \mid q(X)\},$$

so that, among individuals assigned the same calibrated value, the average outcome is close to that value. Under squared loss, the optimal score-only transformation of the original prediction is

$$q_0^{\mathrm{score}}(X) := \mathbb{E}_{P_0}\{Y \mid m(X)\}.$$

Calibration can therefore be viewed as approximating this conditional mean without retraining the underlying prediction model.

We study a simple post-hoc strategy: calibrate $m(X)$ on the labeled sample, then average the calibrated predictions over the pooled labeled and unlabeled samples. We call the resulting approach *calibrated prediction-powered inference*. The estimator has a simple plug-in form and an exact AIPW representation. Different calibration classes yield different estimators, including the linear and isotonic procedures studied below.

3.1 General estimator and exact AIPW representation

Let $m : \mathcal{X} \to \mathbb{R}$ be a black-box score trained on external data or on a separate training sample, and write $T = m(X)$. We estimate a one-dimensional calibration map $g_n : \mathbb{R} \to \mathbb{R}$ from the labeled sample and define the calibrated score

$$q_n := g_n \circ m.$$

Thus $q_n(X)$ is a post-processed version of the original score whose values are better aligned with the observed outcomes.

Let $\mathcal{H}$ be a class of post-processing maps $g : \mathbb{R} \to \mathbb{R}$. We fit g_n by least squares on the labeled sample:

$$g_n \in \arg\min_{g \in \mathcal{H}} \mathbb{P}_n^L\big[(Y - g(T))^2\big]. \tag{2}$$

The calibrated plug-in estimator averages the resulting predictions over all available covariates:

$$\widehat{\psi} := \mathbb{P}_M^X q_n = \frac{1}{M}\Big\{\sum_{i=1}^{n} q_n(X_i) + \sum_{j=1}^{N} q_n(\widetilde{X}_j)\Big\}. \tag{3}$$

If $\mathcal{H}$ is closed under additive constants, optimality of g_n with respect to constant shifts gives

$$\mathbb{P}_n^L\{Y - q_n(X)\} = 0.$$

Consequently, the calibrated plug-in estimator is exactly the AIPW estimator with regression adjustment q_n:

$$\widehat{\psi}_{\mathrm{cal}} = \mathbb{P}_M^X q_n + \mathbb{P}_n^L\{Y - q_n(X)\}. \tag{4}$$

The plug-in estimator need not be unbiased in finite samples because calibration uses the same labeled outcomes. Nonetheless, we show under mild conditions that this bias is asymptotically negligible. Exact finite-sample unbiasedness can be recovered using cross-fitted or leave-one-out AIPW estimators (Appendix D.1).

Why calibration can improve efficiency. For a fixed score q, the asymptotic variance of $\sqrt{n}\{\widehat{\psi}(q) - \psi_0\}$ can also be written as

$$V(q) = \rho_0 \mathrm{Var}(Y) + (1 - \rho_0)\mathrm{Var}\{Y - q(X)\}.$$

Let g_0 minimize $\mathbb{E}[\{Y - g(T)\}^2]$ over $g \in \mathcal{H}$. Since adding a constant to a score does not change $V(q)$, $q_0 = g_0 \circ m$ also minimizes the asymptotic variance over the same class of transformed scores. If the identity map belongs to $\mathcal{H}$, then $V(q_0) \leq V(m)$. If $\mathcal{H}$ contains $t \mapsto \mathbb{E}(Y \mid T = t)$, then $q_0(X) = \mathbb{E}(Y \mid T)$, and AIPW with this score attains the semisupervised efficiency bound based on T rather than the full covariate vector X.

3.2 Isotonic-calibrated PPI

Our main flexible procedure is isotonic calibration, a nonparametric post-hoc calibration method [113, 114, 65]. We consider nondecreasing transformations of the original score, which preserve its ordering up to ties while allowing the predicted values to better match the observed outcomes. This class also includes the identity map, so the original score remains a candidate.

Let $\mathcal{F}_{\mathrm{iso}}$ be the class of nondecreasing maps. Writing $T_i = m(X_i)$, we fit

$$\hat{g}_n^{\mathrm{iso}} \in \arg\min_{g \in \mathcal{F}_{\mathrm{iso}}} \sum_{i=1}^{n} \{Y_i - g(T_i)\}^2, \qquad q_n^{\mathrm{iso}}(X) := \hat{g}_n^{\mathrm{iso}}\{m(X)\},$$

and define

$$\widehat{\psi}_{\mathrm{iso}} := \mathbb{P}_M^X q_n^{\mathrm{iso}}.$$

Isotonic regression requires no tuning parameters and is widely available in standard software. For new score values, we interpolate linearly between the fitted values at distinct observed scores and use the endpoint values outside the observed range.

Proposition 1 (Empirical calibration and risk minimization)**.** *For every measurable $h : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{P}_n^L\big[h(q_n^{\mathrm{iso}}(X))\{Y - q_n^{\mathrm{iso}}(X)\}\big] = 0. \tag{5}$$

Moreover, for every $\theta \in \mathcal{F}_{\mathrm{iso}}$,

$$\mathbb{P}_n^L\big[(Y - q_n^{\mathrm{iso}}(X))^2\big] \leq \mathbb{P}_n^L\big[(Y - \theta(m(X)))^2\big].$$

In particular, isotonic calibration fits no worse on the labeled sample than the original score m.

The orthogonality follows from the blockwise structure of isotonic regression. The fitted score is constant on the blocks determined by the pooled-adjacent-violators algorithm [6], and each fitted value equals the average outcome within its block [98]. Since any function of the fitted score is also constant within these blocks, the residuals are orthogonal to every such function. Histogram binning, scalable monotone implementations, and cross-fitted variants are discussed in Appendices D and O.

3.3 Linearly calibrated PPI

When the labeled sample is small, affine calibration provides a simpler alternative to isotonic regression. It adjusts the intercept and slope of the original score but cannot correct nonlinear miscalibration. We fit

$$(\hat{a}_{\mathrm{lin}}, \hat{b}_{\mathrm{lin}}) \in \arg\min_{a,b \in \mathbb{R}} \sum_{i=1}^{n} \{Y_i - a\, m(X_i) - b\}^2, \qquad q_n^{\mathrm{lin}}(X) := \hat{a}_{\mathrm{lin}}\, m(X) + \hat{b}_{\mathrm{lin}}. \tag{6}$$

The corresponding plug-in estimator is

$$\widehat{\psi}_{\mathrm{lin}} := \mathbb{P}_M^X q_n^{\mathrm{lin}}. \tag{7}$$

The least-squares normal equations imply

$$\mathbb{P}_n^L\big[(a + b q_n^{\mathrm{lin}}(X))\{Y - q_n^{\mathrm{lin}}(X)\}\big] = 0$$

for every $a, b \in \mathbb{R}$. Thus, no further affine transformation of the calibrated score can reduce the labeled-sample squared error.

Linear calibration is closely related to PPI++ [2], which chooses a coefficient multiplying the original score to minimize the estimated asymptotic variance. The variance-minimizing coefficient and the least-squares slope have the same population limit. Let $\hat{\lambda}_{++}$ minimize the empirical AIPW variance over $\lambda \in \mathbb{R}$, and write $\widehat{\psi}_{++}(\hat{\lambda}_{++}) := \widehat{\psi}(\hat{\lambda}_{++} m)$.

Proposition 2 (First-order equivalence of PPI++ and linear calibration)**.** *Assume Assumption 1,* $\mathrm{Var}\{m(X)\} > 0$*, and* $\mathbb{E}_{P_0}\{m(X)^2\} < \infty$*. Then*

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = (1-\rho_n)(\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}})\{\mathbb{P}_N^U m - \mathbb{P}_n^L m\}.$$

Moreover, $\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}} = o_p(1)$*, and hence*

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = o_p(n^{-1/2}).$$

Thus, unrestricted PPI++ and linearly calibrated PPI are first-order equivalent. This connects PPI++ to classical prognostic-score adjustment, where a precomputed outcome prediction is used as a covariate in linear regression to improve precision. Linear calibration gives this approach a simple plug-in interpretation. The equivalence need not hold when the PPI++ coefficient is clipped or constrained [2]. Platt scaling, covariate-augmented linear calibration, and cross-fitting are discussed in Appendix D.

4 Theory for isotonic-calibrated PPI

We next develop first-order theory for isotonic-calibrated PPI and establish two "calibeating" guarantees. First, isotonic calibration minimizes squared prediction error and the asymptotic variance of mean estimation among monotone transformations of the original score. Second, further post-processing of the calibrated score cannot yield a nonvanishing improvement in prediction accuracy or first-order efficiency.

4.1 Asymptotic normality, inference, and efficiency

Let g_0^{iso} be the population isotonic calibration map for the fixed score m:

$$g_0^{\mathrm{iso}} \in \arg\min_{g \in \mathcal{F}_{\mathrm{iso}}} \mathbb{E}_{P_0}\big[\{Y - g(T)\}^2\big], \qquad q_0^{\mathrm{iso}} := g_0^{\mathrm{iso}} \circ m. \tag{8}$$

Thus q_0^{iso} is the best monotone post-processing of the original score under squared loss. Since the class is closed under additive constants, $\mathbb{E}_{P_0}[q_0^{\mathrm{iso}}(X)] = \psi_0$.

Assumption 2 (Regularity for isotonic calibration)**.** The following conditions hold.

(i) **Boundedness.** There exists $C_0 < \infty$ such that $|\mathbb{E}_{P_0}[Y \mid m(X)]| \le C_0$ almost surely, and $\sup_x |q_n^{\mathrm{iso}}(x)| = O_p(1)$.

(ii) **Tails.** For $T = m(X)$, there exists $0 < K < \infty$ such that

$$\mathbb{E}_{P_0}\left[\exp\left\{\frac{|Y - \mathbb{E}_{P_0}[Y \mid T]|}{K}\right\} \middle| T\right] \le 2 \qquad \text{almost surely.}$$

Theorem 1 (Asymptotic linearity under the two-sample model)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$\|q_n^{\mathrm{iso}} - q_0^{\mathrm{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Define

$$D^L_{q_0^{\mathrm{iso}},n}(X,Y) := Y - \psi_0 - (1-\rho_n)\{q_0^{\mathrm{iso}}(X) - \psi_0\}, \qquad D^U_{q_0^{\mathrm{iso}},n}(\widetilde{X}) := (1-\rho_n)\{q_0^{\mathrm{iso}}(\widetilde{X}) - \psi_0\}.$$

Then

$$\widehat{\psi}_{\mathrm{iso}} - \psi_0 = (\mathbb{P}_n^L - P_0)D^L_{q_0^{\mathrm{iso}},n} + (\mathbb{P}_N^U - P_{0,X})D^U_{q_0^{\mathrm{iso}},n} + R_{n,N}, \tag{9}$$
$$R_{n,N} = O_p\big((1-\rho_n)n^{-2/3}\big) + O_p\big((1-\rho_n)n^{-1/3}N^{-1/2}\big) = o_p(n^{-1/2}).$$

Consequently, $\widehat{\psi}_{\mathrm{iso}}$ *is asymptotically linear at the* $\sqrt{n}$ *rate with triangular influence pair* $(D^L_{q_0^{\mathrm{iso}},n}, D^U_{q_0^{\mathrm{iso}},n})$.

The asymptotic variance of $\sqrt{n}(\widehat{\psi}_{\text{iso}} - \psi_0)$ is

$$\sigma_0^2 := \text{Var}\big[Y - \psi_0 - (1-\rho_0)\{q_0^{\text{iso}}(X) - \psi_0\}\big] + \rho_0(1-\rho_0)\text{Var}\{q_0^{\text{iso}}(X)\}.$$

The next result establishes asymptotic normality and valid Wald inference using an empirical variance estimator.

Corollary 1 (Asymptotic normality and Wald inference)**.** *Under the conditions of theorem 1, assume that $\sigma_0^2 > 0$. Define the empirical variance estimator*

$$\widehat{\sigma}^2 := \mathbb{P}_n^L\{(\widehat{D}^L - \mathbb{P}_n^L\widehat{D}^L)^2\} + \frac{n}{N}\mathbb{P}_N^U\{(\widehat{D}^U - \mathbb{P}_N^U\widehat{D}^U)^2\},$$

where

$$\widehat{D}_i^L := Y_i - \widehat{\psi}_{\text{iso}} - (1-\rho_n)\{q_n^{\text{iso}}(X_i) - \widehat{\psi}_{\text{iso}}\}, \qquad \widehat{D}_j^U := (1-\rho_n)\{q_n^{\text{iso}}(\widetilde{X}_j) - \widehat{\psi}_{\text{iso}}\}.$$

Then

$$\frac{\sqrt{n}\,(\widehat{\psi}_{\text{iso}} - \psi_0)}{\widehat{\sigma}} \xrightarrow{d} N(0,1).$$

Corollary 2 (Efficiency)**.** *Under the conditions of theorem 1, the following hold.*

(i) If $q_0^{\text{iso}} = \mu_0$ almost surely, then $\widehat{\psi}_{\text{iso}}$ is efficient in the full two-sample model.

(ii) For any nondecreasing $\theta : \mathbb{R} \to \mathbb{R}$ such that $\theta \circ m$ is square-integrable, $\widehat{\psi}_{\text{iso}}$ has asymptotic variance no larger than the AIPW estimator with score $\theta \circ m$.

(iii) If $\mathbb{E}_{P_0}[Y \mid m(X)]$ is nondecreasing in $m(X)$, then

$$q_0^{\text{iso}}(X) = \mathbb{E}_{P_0}[Y \mid m(X)] \quad \textit{almost surely}.$$

In this case, no AIPW estimator based on a square-integrable transformation $\theta \circ m$ has smaller asymptotic variance. Equivalently, $\widehat{\psi}_{\text{iso}}$ is efficient in the reduced two-sample model that observes the scalar score $m(X)$ instead of X.

Taking θ constant in part (ii) shows that isotonic calibration is asymptotically at least as efficient as the labeled mean. The reduction in asymptotic variance is

$$\text{Var}(Y) - \sigma_0^2 = (1-\rho_0)\text{Var}\{q_0^{\text{iso}}(X)\}.$$

This follows from the orthogonality of the population isotonic regression residuals to the calibrated score.

The efficiency guarantee also applies relative to AIPW with the raw score, PPI, and linear calibration with a nonnegative population slope. By the equivalence established in proposition 2, it also applies relative to PPI++ when its limiting scaling coefficient is nonnegative. For $\rho_0 < 1$, full efficiency in the original two-sample model requires $q_0^{\text{iso}} = \mu_0$. However, when $\mathbb{E}_{P_0}[Y \mid m(X)]$ is nondecreasing in $m(X)$, isotonic calibration is first-order optimal among all square-integrable score-only adjustments.

When $N/n \to \infty$, the limiting labeled fraction is zero, and the asymptotic variance reduces to

$$\sigma_0^2 = \text{Var}\{Y - q_0^{\text{iso}}(X)\}.$$

If the conditional mean of Y given $m(X)$ is nondecreasing in the score, this equals $\mathbb{E}[\text{Var}\{Y \mid m(X)\}]$. In this regime, the unlabeled contribution is negligible at the $\sqrt{n}$ scale, and the labeled residuals determine first-order uncertainty.

4.2 No further first-order gain from post-hoc calibration

The previous result compares isotonic calibration with monotone transformations of the original score. We now consider whether further post-processing of the fitted isotonic score can improve prediction or inference. Among all predictors based only on $q_n^{\text{iso}}(X)$, the optimal squared-error predictor is

$$q_{0|n}(X) := \mathbb{E}_{P_0}\{Y \mid q_n^{\text{iso}}(X)\},$$

where the conditional expectation is evaluated for a new observation from P_0, with the fitted isotonic map held fixed.

Assumption 3 (Bounded-variation score regression)**.** The map $t \mapsto \mathbb{E}_{P_0}[Y \mid m(X) = t]$ has bounded variation on the support of $m(X)$.

This assumption allows the conditional mean to be nonmonotone and discontinuous. In particular, it does not require differentiability or Lipschitz continuity.

Theorem 2 (No further first-order gain from post-hoc calibration)**.** *Suppose Assumptions 1 to 3 hold. Let*

$$\widehat{\psi}_{\text{eff}} := \widehat{\psi}(q_{0|n}).$$

Then

$$\|q_{0|n} - q_n^{\text{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}), \qquad \widehat{\psi}_{\text{eff}} - \widehat{\psi}_{\text{iso}} = o_p(n^{-1/2}).$$

Thus $\widehat{\psi}_{\text{iso}}$ and $\widehat{\psi}_{\text{eff}}$ have the same asymptotic variance.

The first conclusion shows that further post-processing of the isotonic score can reduce population squared prediction error by at most $O_p(n^{-2/3})$. The second shows that even replacing the calibrated score by the optimal score-only predictor does not change the estimator at the first order. Thus no further post-processing of the fitted isotonic score can improve first-order efficiency.

Unlike corollary 2, this result allows arbitrary measurable transformations of the fitted isotonic score, rather than only monotone transformations of the original score.

5 Experiments

We compare the labeled-only mean, AIPW, PPI, PPI++, and AIPW-EM with four calibration-based methods: `LinearCal`, `MonoSpline`, `IsoCal`, and `AutoCal`. LinearCal fits an affine transformation, MonoSpline fits a smooth nondecreasing transformation, and IsoCal fits a nondecreasing transformation. AutoCal selects among AIPW and these calibration methods using cross-validated estimates of the AIPW variance. In the real-data and LLM benchmarks, LinearCal clips its predictions to the labeled outcome range, while IsoCal groups observations into bins of at least 10 before fitting isotonic regression. Appendix N gives the implementation details.

We first examine score miscalibration and small labeled samples in simulations, then evaluate the methods on public PPI benchmarks and LLM evaluation tasks. We report bias, Wald-interval coverage, MSE, and Monte Carlo standard deviation. Empirical relative efficiency is the ratio of the reference estimator's Monte Carlo variance to that of the estimator being evaluated; each figure caption specifies the reference estimator.

Simulation study We generate $S \sim N(0,1)$ and $Y \mid S \sim \text{Bernoulli}\{\mu_0(S)\}$, where $\mu_0(S) = \sigma(5S)$ and $\sigma(t) = (1+e^{-t})^{-1}$. The prediction score is

$$m(X) = C[-0.15 + 0.75\,\sigma\{3\,\text{logit}\{\mu_0(S)\} - 0.5\}]\,,$$

where $C(t) = \min\{0.99, \max(0.01, t)\}$. This transformation preserves the ordering of the conditional mean, apart from ties induced by clipping, but introduces miscalibration on the probability scale. We consider $n \in \{50, 100, 200, 400, 800, 1200, 2400\}$, with $N/n \in \{1, 16\}$, and use 500 Monte Carlo replications per setting.

At $n = 400$ and $N/n = 16$, `LinearCal`, `MonoSpline`, and `IsoCal` have RMSEs between 0.0164 and 0.0166, compared with 0.0172 for PPI and 0.0175 for AIPW. Their Wald-interval coverage ranges from 0.910 to 0.932.

The PPI++ implementation restricts the scaling coefficient to $[0, 1]$ in the original PPI parameterization, equivalent to $[0, (1-\rho_n)^{-1}]$ in the AIPW parameterization. AIPW-EM minimizes the same empirical variance criterion without this constraint and performs similarly to `LinearCal`. Averaged across sample sizes with $N/n = 16$, RMSE is approximately 0.0208 for PPI++, compared with 0.0196 for AIPW-EM and `LinearCal`.

Small-sample comparisons We also examine smaller labeled samples under the same binary-outcome model, with mild and strong score miscalibration as specified in Appendix N.1. We set $N = 16n$ for $n \in \{20, 50, 100, 200, 400, 800, 1200\}$ and use 1,000 replications per setting. The

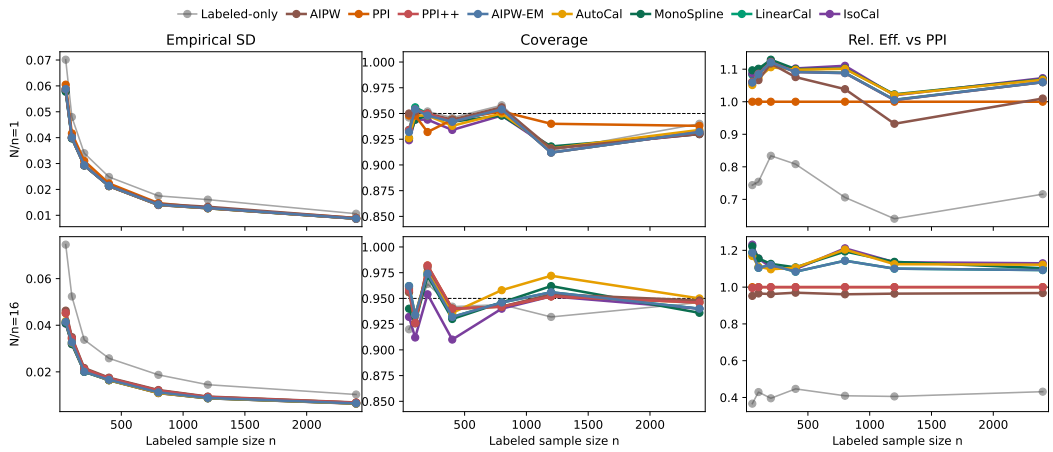

Figure 1: Synthetic simulation study with a poorly calibrated score. The top row shows $N/n = 1$, and the bottom row shows $N/n = 16$. Columns report Monte Carlo standard deviation, coverage, and relative efficiency versus PPI.

labeled mean, PPI, AIPW, PPI++, and the three calibration methods are evaluated on the same samples, together with the leave-one-out linear and isotonic estimators in (14).

Under strong miscalibration, linear, smooth monotone, and isotonic calibration each have lower RMSE than PPI++ at all seven sample sizes. At $n = 20$, isotonic coverage ranges from 0.814 to 0.871 across the two settings; leave-one-out isotonic calibration increases coverage to 0.905–0.925, with wider intervals and larger RMSE. Across all 14 settings, leave-one-out calibration reduces linear-calibration RMSE by at most 1.99%, but reduces isotonic RMSE in only one setting. Comparisons with random-forest and neural-network implementations of RePPI are reported in Appendix N.1.

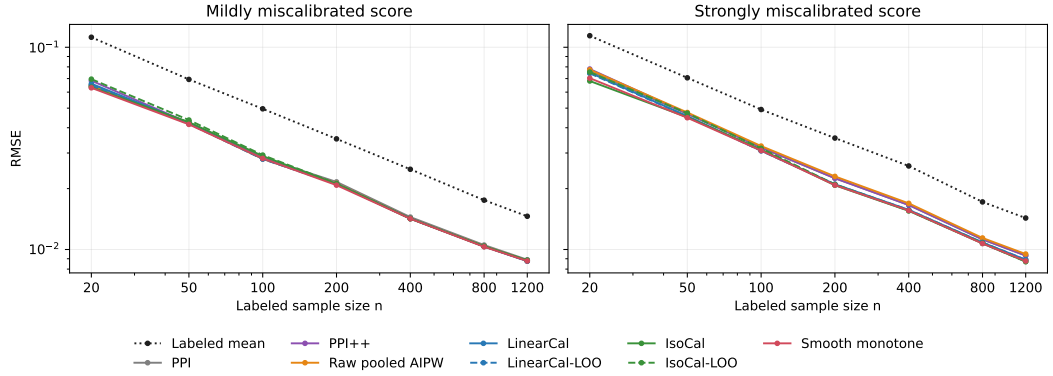

Figure 2: RMSE across labeled sample sizes under mild and strong score miscalibration, with $N = 16n$ and 1,000 replications per setting. Solid and dashed curves distinguish calibration on the labeled sample from leave-one-out calibration.

Reproduced PPI benchmarks We next evaluate the methods on the standard PPI mean-estimation benchmarks of [1], where any score miscalibration arises from the public prediction scores rather than being introduced by design. Using the public `ppi_py` `forest`, `galaxies`, and `census_income` datasets, we rerandomize the labeled/unlabeled split 500 times at each labeled-sample size and evaluate all estimators on the same splits.

AIPW is generally the strongest raw-score baseline, while PPI is competitive mainly when the unlabeled sample is extremely large. The benefits of calibration vary across datasets. On `forest` and `galaxies`, the best-performing score-based methods have similar MSE. On `census_income`, the

score is sufficiently mis-scaled that `LinearCal` has the lowest average MSE. Across the 30 combinations of dataset and labeled-sample size, unclipped AIPW-EM has lower MSE than clipped PPI++ in 29 combinations and is nearly tied with `LinearCal`. This suggests that much of the observed difference between PPI++ and empirical efficiency maximization is attributable to clipping. Appendix N gives the dataset-by-regime tables and full calibration comparisons.

LLM evaluation benchmark We also study LLM evaluation, where evaluator scores serve as proxies for human preferences or response correctness. The first track uses PPE Human Preference V1 [26]; the second uses a PPE correctness suite built from MMLU-Pro, MATH, GPQA, IFEval, and MBPP+ [26, 110, 39, 73, 117, 59]. In PPE Human, the estimand for each target model is its human win rate, counting ties as half-wins. In PPE Correctness, each observation is a pair containing one correct and one incorrect response from the same model, and the estimand is the probability that the first response is correct. We use four public precomputed evaluator-score sources on the PPE tracks: Skywork-Reward-Llama-3.1-8B, ArmoRM-Llama3-8B-v0.1, Athene-RM-8B, and Starling-RM-7B-alpha. We use 200 random labeled/unlabeled splits for each combination of evaluator, benchmark, and target model.

On PPE Human at $n = 100$, raw PPI and AIPW have MSEs of 0.495 and 0.440, compared with 0.001425 for labeled-only estimation and 0.001391 for `LinearCal`. On PPE Correctness, the corresponding MSEs are 0.534, 0.353, 0.002010, and 0.001813. The large errors of raw PPI and AIPW reflect the poor numerical scale of the uncalibrated evaluator scores. Relative to labeled-only estimation, `LinearCal` reduces average MSE by 2.4% on PPE Human and 9.8% on PPE Correctness. Its nominal 90% Wald intervals have empirical coverage of 0.909 and 0.931, respectively. Table 1 summarizes the results; evaluator-specific comparisons and model-ranking results appear in Appendices N.3 and N.4.

The ArmoRM results on MATH provide a separate comparison of score ordering and calibration. Empirical correctness increases across twelve score bins, from 0.078 in the lowest bin to 0.920 in the highest. Platt calibration fitted on 100 labels reduces prediction MSE from 0.4683 to 0.2020 on held-out observations. Appendix N.2 gives the calibration plot and details.

6 Conclusion

We propose calibrated PPI, which estimates a population mean by calibrating black-box predictions using labeled observations and averaging the calibrated predictions over the pooled covariate sample. When the calibration class is closed under additive constants, the resulting plug-in estimator has an exact AIPW representation. We show that linear calibration is first-order equivalent to unconstrained PPI++. For isotonic calibration, we establish asymptotic normality and valid Wald inference without sample splitting, optimality among nondecreasing transformations of the original score, and no further first-order efficiency gain from post-processing the fitted score. Simulations and benchmarks show that calibration can substantially improve precision when the original predictions are poorly calibrated, without retraining the underlying prediction model.

Acknowledgments and Disclosure of Funding

The authors have no funding to disclose. The authors declare no competing interests.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[2] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

[3] Daniele Ballinari and Nora Bearth. Improving the finite sample performance of double/debiased machine learning with propensity score calibration. *arXiv preprint arXiv:2409.04874*, 2024.

[4] Laura B Balzer, Erica Cai, Lucas Godoy Garraza, and Pracheta Amaranath. Adaptive selection of the optimal strategy to improve precision and power in randomized trials. *Biometrics*, 80(1):ujad034, 2024.

[5] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[6] Richard E. Barlow and Hugh D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

[7] Aurélien F. Bibaut and Mark J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019.

[8] Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

[9] David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L. Ogburn. Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf019, 2025.

[10] Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620, 1976.

[11] Ambarish Chattopadhyay and José R. Zubizarreta. On the implied weights of linear regression for causal inference. *Biometrika*, 110(3):615–629, 2023.

[12] Ambarish Chattopadhyay, Christopher H Hase, and José R Zubizarreta. Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, 39(24):3227–3254, 2020.

[13] Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased machine learning without sample-splitting for stable estimators. In *Advances in Neural Information Processing Systems*, volume 35, pages 30925–30937, 2022.

[14] Song Xi Chen, Denis HY Leung, and Jing Qin. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):803–823, 2008.

[15] Yi-Hau Chen. Miscellanea. a robust imputation method for surrogate outcome data. *Biometrika*, 87(3):711–716, 2000.

[16] Yuxi Chen, Edward H Kennedy, and Sivaraman Balakrishnan. On the equivalence between neyman orthogonality and pathwise differentiability. *arXiv preprint arXiv:2603.15817*, 2026.

[17] David Cheng, Ashwin N Ananthakrishnan, and Tianxi Cai. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 77(2):413–423, 2021.

[18] 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.

[19] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

[20] Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

[21] David R Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4): 562–565, 1958.

[22] Shachi Deshpande and Volodymyr Kuleshov. Calibrated and conformal propensity scores for causal effect estimation. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2023.

[23] Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

[24] Peng Ding and Fan Li. Causal inference: A missing data perspective, 2018. URL https://arxiv.org/abs/1712.06170.

[25] Dean P Foster and Sergiu Hart. "calibeating": Beating forecasters at their own game. *Theoretical Economics*, 18(4):1441–1474, 2023.

[26] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cbttLtO94Q.

[27] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

[28] Ellen Graham, Marco Carone, and Andrea Rotnitzky. Towards a unified theory for semiparametric data fusion with individual-level data. *arXiv preprint arXiv:2409.09973*, 2024.

[29] Susan Gruber and Mark J Van Der Laan. Targeted maximum likelihood estimation: A gentle introduction. 2009.

[30] Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International conference on machine learning*, pages 3942–3952. PMLR, 2021.

[31] Chirag Gupta and Aaditya Ramdas. Online platt scaling with calibeating. In *International Conference on Machine Learning*, pages 12182–12204. PMLR, 2023.

[32] Rom Gutman, Ehud Karavani, and Yishai Shimoni. Propensity score models are better when post-calibrated. *arXiv preprint arXiv:2211.01221*, 2022.

[33] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[34] Peisong Han. A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics & Probability Letters*, 82(12):2221–2228, 2012.

[35] Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

[36] Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, volume 473, pages 474–482. North-Holland Amsterdam, 1979.

[37] Nima S. Hejazi and Mark J. van der Laan. Revisiting the propensity score's central role: Towards bridging balance and efficiency in the era of causal machine learning. *Observational Studies*, 9(1):23–34, 2023.

[38] Nima S Hejazi, Mark J van der Laan, Holly E Janes, Peter B Gilbert, and David C Benkeser. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*, 77(4):1241–1253, 2021.

[39] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

[40] Katherine Hoffman. An illustrated guide to tmle, part i: Introduction and motivation, December 2020. URL https://www.khstats.com/blog/tmle/tutorial.

[41] Emilie Højbjerre-Frandsen and Alejandro Schuler. "within-trial" prognostic score adjustment is targeted maximum likelihood estimation. *Pharmaceutical Statistics*, 25(2):e70080, 2026.

[42] Emilie Højbjerre-Frandsen, Mark J van der Laan, and Alejandro Schuler. Powering rcts for marginal effects with glms using prognostic score adjustment. *arXiv preprint arXiv:2503.22284*, 2025.

[43] Chuan Hong, Katherine P. Liao, and Tianxi Cai. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics*, 75(1):78–89, 2019. doi: 10.1111/biom.12971.

[44] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B*, 76(1):243–263, 2014.

[45] Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *arXiv preprint arXiv:2501.09731*, 2025.

[46] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

[47] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025.

[48] Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

[49] Nicolas S Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 3(5):18, 2011.

[50] Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335): 1–72, 2023.

[51] Se Yoon Lee and Jae-Kwang Kim. Mec: Machine-learning-assisted generalized entropy calibration for semi-supervised mean estimation. *arXiv preprint arXiv:2604.05446*, 2026.

[52] Samuel D. Lendle, Bruce Fireman, and Mark J. van der Laan. Balancing score adjusted targeted minimum loss-based estimation. *Journal of Causal Inference*, 3(2):139–155, 2015.

[53] Yan Leng and Drew Dimmery. Calibration of heterogeneous treatment effects in random experiments. *Available at SSRN 3875850*, 2021.

[54] Boris Yakovlevich Levit. On efficiency of a class of non-parametric estimates. *Teoriya Veroyatnostei i ee Primeneniya*, 20(4):738–754, 1975.

[55] Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054, 2023.

[56] Sijia Li, Peter B Gilbert, Rui Duan, and Alex Luedtke. Data fusion using weakly aligned sources. *Journal of the American Statistical Association*, 120(552):2569–2579, 2025.

[57] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer, 1977.

[58] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, pages 295–318, 2013.

[59] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

[60] Jacob A Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER, 1969.

[61] Kelly L Moore and Mark J van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1): 39–64, 2009.

[62] Kelly L Moore and Mark J van der Laan. Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of biopharmaceutical statistics*, 19 (6):1099–1131, 2009.

[63] Kelly L Moore, Romain Neugebauer, Thamban Valappil, and Mark J van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in medicine*, 30(19):2389–2408, 2011.

[64] Reagan Mozer. Ppi is the difference estimator: Recognizing the survey sampling roots of prediction-powered inference. *arXiv preprint arXiv:2603.19160*, 2026.

[65] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

[66] Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.

[67] Margaret Sullivan Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2):355–365, 1992.

[68] Margaret Sullivan Pepe, Marie Reilly, and Thomas R Fleming. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42(1-2):137–160, 1994.

[69] J Pfanzagl and W Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

[70] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[71] Pierre-Emmanuel Poulet, Maylis Tran, Sophie Tezenas du Montcel, Bruno Dubois, Stanley Durrleman, Bruno Jedynak, and Alzheimer's Disease Neuroimaging Initiative. Prediction-powered inference for clinical trials: application to linear covariate adjustment. *BMC Medical Research Methodology*, 25(1):204, 2025.

[72] Sky Qiu, Susan Gruber, Pamela A Shaw, Brian D Williamson, and Mark J van der Laan. Efficient targeted maximum likelihood estimators for two-phase design problems. *arXiv preprint arXiv:2602.24131*, 2026.

[73] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

[74] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[75] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[76] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[77] Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1):17, 2011.

[78] Michael Rosenblum and Mark J van der Laan. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945, 2009.

[79] Rachael K Ross, Lina M Montoya, Dana E Goin, Iván Díaz, and Audrey Renson. Constructing targeted minimum loss/maximum likelihood estimators: a simple illustration to build intuition. *American journal of epidemiology*, page kwaf261, 2025.

[80] Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

[81] Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

[82] Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. A note on the relation between one-step, outcome regression and ipw-type estimators of parameters with the mixed bias property. *arXiv preprint arXiv:2509.22452*, 2025.

[83] Daniel B Rubin and Mark J van der Laan. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

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

[85] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[86] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[87] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[88] Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

[89] Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, Critical Path for Alzheimer's Disease, Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022.

[90] Andrew J Scott and D Holt. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380):848–854, 1982.

[91] Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33 (2):184, 2018.

[92] Yilin Song, Dan M Kluger, Harsh Parikh, and Tian Gu. Demystifying prediction powered inference. *arXiv preprint arXiv:2601.20819*, 2026.

[93] Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

[94] Lars van der Laan and Ahmed Alaa. Generalized venn and venn-abers calibration with applications in conformal prediction. *arXiv preprint arXiv:2502.05676*, 2025.

[95] Lars van der Laan and Ahmed M Alaa. Self-calibrating conformal prediction. *Advances in Neural Information Processing Systems*, 37:107138–107170, 2024.

[96] Lars van der Laan and Nathan Kallus. Bellman calibration for v-learning in offline reinforcement learning. *arXiv preprint arXiv:2512.23694*, 2025.

[97] Lars van der Laan, Marco Carone, Alex Luedtke, and Mark van der Laan. Adaptive debiased machine learning using data-driven model selection techniques. *arXiv preprint arXiv:2307.12544*, 2023.

[98] Lars van der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *International Conference on Machine Learning*, pages 34831–34854, 2023.

[99] Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration. *arXiv preprint arXiv:2411.06342*, 2024.

[100] Lars van der Laan, Alex Luedtke, and Marco Carone. Doubly robust inference via calibration. *arXiv preprint arXiv:2411.02771*, 2024.

[101] Lars van der Laan, David Hubbard, Allen Tran, Nathan Kallus, and Aurélien Bibaut. Semiparametric double reinforcement learning with applications to long-term causal inference. *arXiv preprint arXiv:2501.06926*, 2025.

[102] Mark van der Laan. Cv-tmle and double machine learning, December 2019. URL https://vanderlaan-lab.org/2019/12/24/cv-tmle-and-double-machine-learning/.

[103] Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

[104] Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

[105] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. 2006.

[106] Adriaan W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[107] Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

[108] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[109] Volodimir G Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2): 245–277, 1992.

[110] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

[111] J Whitehouse, C Jung, V Syrgkanis, B Wilder, and ZS Wu. Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933*, 2024.

[112] Zichun Xu, Daniela Witten, and Ali Shojaie. A unified framework for semiparametrically efficient semi-supervised learning. *arXiv preprint arXiv:2502.17741*, 2025.

[113] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.

[114] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[115] Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N. Ananthakrishnan, Zongqi Xia, Stanley Y. Shaw, Vivian Gainer, Victor Castro, Nicholas Link, Jacqueline Honerlaw, Sicong Huang, David Gagnon, Elizabeth W. Karlson, Robert M. Plenge, Peter Szolovits, Guergana Savova, Susanne Churchill, Christopher O'Donnell, Shawn N. Murphy, J. Michael Gaziano, Isaac Kohane, Tianxi Cai, and Katherine P. Liao. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature Protocols*, 14(12):3426–3444, 2019. doi: 10.1038/s41596-019-0227-6.

[116] Wenjing Zheng and Mark J Van Der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. 2010.

[117] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

[118] Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

[119] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A LLM evaluation results

For each benchmark, evaluator, and target model, MSE/PPI is the estimator's Monte Carlo MSE divided by that of PPI. Implied label savings are

$$\max\left\{0,\, 1 - \frac{\widehat{\mathrm{Var}}(\widehat{\psi})}{\widehat{\mathrm{Var}}(\overline{Y}_L)}\right\},$$

where the variances are computed over repeated splits and $\overline{Y}_L$ is the labeled mean. We set negative savings to zero before averaging. Table entries are averaged first over target models within each benchmark, then over benchmarks within each evaluator, and then over evaluators.

Appendix Contents

The appendix is ordered from background and practical guidance to extensions, proofs, empirical details, and the implementation sketch.

A. LLM evaluation results
B. Extended related work
C. Semiparametric background
D. Additional remarks
E. Calibrated DML for randomized trials
F. Moment equations under missing at random
G. Pooled i.i.d. missing-data formulation
H. AIPW comparison identity
I. PPI++ and linear calibration
J. Proof of the AIPW influence-pair characterization
K. Supporting lemmas for calibration
L. Proof of asymptotic linearity

Track	n	Estimator	MSE / PPI	Label savings	Coverage
PPE Correctness	100	Labeled-only	0.346	0.000	0.927
PPE Correctness	100	PPI	1.000	0.010	0.904
PPE Correctness	100	AIPW	0.774	0.009	0.906
PPE Correctness	100	LinearCal	0.305	0.101	0.931
PPE Correctness	100	AutoCal	0.310	0.094	0.929
PPE Correctness	400	Labeled-only	0.299	0.000	0.988
PPE Correctness	400	PPI	1.000	0.020	0.924
PPE Correctness	400	AIPW	0.405	0.009	0.941
PPE Correctness	400	LinearCal	0.260	0.105	0.989
PPE Correctness	400	AutoCal	0.263	0.102	0.988
PPE Human Preference	100	Labeled-only	0.303	0.000	0.911
PPE Human Preference	100	PPI	1.000	0.004	0.897
PPE Human Preference	100	AIPW	0.919	0.004	0.896
PPE Human Preference	100	LinearCal	0.295	0.031	0.909
PPE Human Preference	100	AutoCal	0.297	0.028	0.908
PPE Human Preference	400	Labeled-only	0.289	0.000	0.945
PPE Human Preference	400	PPI	1.000	0.005	0.915
PPE Human Preference	400	AIPW	0.703	0.004	0.917
PPE Human Preference	400	LinearCal	0.282	0.028	0.944
PPE Human Preference	400	AutoCal	0.282	0.027	0.944

Table 1: PPE LLM-evaluation summary at $n \in \{100, 400\}$, macro-averaged across the public evaluator models. MSE/PPI reports MSE normalized to raw PPI; implied label savings is relative to labeled-only; coverage is for nominal 90% Wald intervals.

M. Proof of post-processing equivalence and reduced-model efficiency

N. Additional empirical details

O. Implementation sketch

B Extended related work

Although PPI is often presented as a recent development, the underlying estimation problem and many of its estimators are special cases of a broader literature on semiparametric debiasing and regression adjustment, including work on missing data, causal inference, and debiased machine learning.

Semiparametric statistics and missing data. Our problem sits squarely in the missing-data and semiparametric literatures. Rubin [84] formalized missing at random as a key identification condition for recovering full-data targets from partially observed outcomes. Semiparametric efficiency theory then characterized efficient influence functions, regular asymptotically linear estimators, and one-step debiased procedures [54, 69, 36, 48, 8]. In missing-data and causal-inference problems, these ideas led to influence-function-based estimating equations [75, 76, 74, 103, 93], often with doubly robust structure [5]. Accordingly, the efficiency theory for missing-data models under missingness or coarsening at random is well understood, and the class of regular asymptotically linear estimators in these models can be characterized through their influence functions [75, 76, 74, 103, 93].

Debiased machine learning and targeted learning. A large body of work combines the semiparametric framework above with machine-learning-based nuisance estimation, often under the label of debiased machine learning. Flexible learning methods within estimating equations and AIPW-style procedures already appear in van der Laan and Robins [103]. Targeted minimum loss estimation (TMLE), or targeted learning, updates an initial nuisance estimator in a parameter-targeted way so that the resulting plug-in estimator is debiased; in the present setting, the main nuisance is the outcome regression [105, 104, 29, 40, 79]. These approaches are often combined with sample splitting and cross-fitting to accommodate generic machine-learning estimators [88, 48, 116, 104, 18]. Double machine learning emphasizes similar ingredients—orthogonality, cross-fitting, and influence-

function expansions—for inference with modern nuisance estimators [18, 20], and can be understood within the same semiparametric framework [102, 23, 16]. For sufficiently stable estimators, cross-fitting can itself be relaxed [13]. Closely related work studies balancing-weight estimators [33, 44, 119, 52, 12, 37], which often admit equivalent regression-adjustment or augmented representations [11, 9, 82]. In completely missing-at-random settings, including semisupervised learning and randomized trials, these approaches allow flexible black-box nuisance estimation while still supporting valid inference under appropriate conditions [61, 62, 78, 42, 41].

Semisupervised inference. Related work studies semisupervised or surrogate-assisted estimation in settings where the primary outcome is observed only on a limited subset, while auxiliary outcomes or surrogate measurements are available more broadly [67, 68, 15, 14, 17, 45, 47]. Two-sample semisupervised inference is a special case of two-stage sampling designs [90, 77, 38, 72] and, more generally, of data fusion, for which a substantial semiparametric efficiency literature has been developed [55, 56, 28, 112]. More recent papers have used the term *prediction-powered inference* (PPI) for the semisupervised mean-estimation setting in which a black-box predictor is combined with a small labeled sample and a large unlabeled sample [1, 2, 118, 112, 92, 71, 51]. These papers also helped popularize such methods for a broader machine-learning audience and provided accessible software. Closely related questions have long been studied in randomized trials, where prognostic scores or other regression adjustments learned from larger historical or auxiliary data sources are used to improve precision in smaller trials [35, 78, 61–63, 89, 4, 42]. Methodologically, however, PPI estimators are best understood as instances or variants of existing semiparametric procedures. For example, PPI++ is an AIPW estimator [75, 103] obtained by empirical efficiency maximization [83] over a scaling parameter, and cross-PPI [118] is simply AIPW with cross-fitting. MEC [51] instead uses generalized entropy weights to balance the prediction score, which is closer to the balancing-weight representation of our linearly calibrated estimator. Our linearly calibrated estimator emphasizes the corresponding regression-adjustment view, implemented by ordinary linear regression of Y on the score. More fundamentally, the ability to combine black-box prediction models with valid inference is a consequence of the double robustness of AIPW when the missingness probability is known, and is not specific to semisupervised mean estimation [5, 91, 81].

From missing data to causal inference. From the perspective of estimation and efficiency theory, semisupervised mean estimation with missing-at-random outcomes is equivalent to randomized experiments, controlled trials, and survey-sampling designs once the missingness indicator is reinterpreted as a treatment or sampling indicator, so that the relevant potential outcomes are viewed as missing [85, 103, 24, 64]. This equivalence allows tools from covariate adjustment, causal inference, and model-assisted estimation to be transferred directly to the semisupervised setting; see, for example, Mozer [64] for results relating prediction-powered inference to model-assisted survey-sampling estimators. The two-sample formulation we study is also equivalent to a one-sample i.i.d. model [55]. In particular, our methods apply to inference on the counterfactual mean $\mathbb{E}[Y(1)]$ in a randomized trial with binary treatment, where treated units form the labeled sample and control units contribute covariate information only; see Appendix E.

Calibration. A final relevant thread concerns calibration. Classical post-processing methods such as Platt scaling [70, 21, 31], linear calibration [60, 19, 53], isotonic regression [114, 65], and histogram binning [113, 30] improve predictive reliability without retraining the underlying model. The idea has deeper roots in the forecasting literature [60, 57, 109, 108, 27, 49]; see also the historical discussion in Lee et al. [50]. More recent work studies calibration under broader losses and prediction tasks [46, 66, 98, 111, 94], as well as applications to conformal prediction [95, 94], importance-weight stabilization [32, 22, 3, 99], treatment effect estimation [19, 98, 111], and value prediction in offline reinforcement learning [96]. Distribution-free guarantees for histogram binning are developed by Gupta and Ramdas [30], while van der Laan et al. [98], van der Laan and Alaa [94] establish asymptotic distribution-free guarantees for isotonic calibration. Most closely related to our work is recent research connecting calibration to debiased estimation of statistical functionals [100, 97, 101]. In particular, van der Laan et al. [100] studies calibrated debiased machine learning for doubly robust inference, which in our setting reduces to calibrating the outcome regression and exploiting the resulting plug-in representation, while van der Laan et al. [101] studies calibrated plug-in regression estimators in reinforcement learning. More broadly, van der Laan et al. [97] shows how plug-in calibration can support regression and treatment effect estimation, including superefficient inference through dimension reduction.

C Semiparametric background

C.1 Review of AIPW, PPI, and semiparametric efficiency

The AIPW class. We begin with the augmented inverse-probability weighted (AIPW) estimator in the unrestricted two-sample model [75]. We view AIPW not as a single estimator, but as a family of estimators indexed by a regression-adjustment score $q : \mathcal{X} \to \mathbb{R}$ with finite second moment. For any such q, define

$$\widehat{\psi}(q) := \mathbb{P}_M^X q + \mathbb{P}_n^L\{Y - q(X)\}. \tag{10}$$

This estimator combines a plug-in term that imputes Y from $q(X)$ with a residual correction from the labeled sample. For any fixed q, $\widehat{\psi}(q)$ is unbiased in finite samples for $\psi_0 = \mathbb{E}_{P_0}[Y]$. In particular, $q(X) = 0$ yields the empirical mean estimator $\mathbb{P}_n^L\{Y\}$. Thus, the choice of q affects efficiency, through the estimator's variance, but not validity, in the sense that the resulting confidence intervals remain valid. This makes AIPW a "safe" estimator [112]. This property can be viewed as a particularly simple instance of double robustness[2] [5]. The same basic principle underlies the use of flexible machine learning for covariate adjustment in randomized trials while retaining valid inference [61, 62, 89, 42].

Asymptotic linearity and efficiency. We say that $\widehat{\psi}$ is asymptotically linear with triangular influence pair (D_n^L, D_n^U) if

$$\widehat{\psi} - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D_n^L(X_i, Y_i) + \frac{1}{N}\sum_{j=1}^{N} D_n^U(\widetilde{X}_j) + o_p(n^{-1/2}),$$

where $\mathbb{E}_{P_0}[D_n^L(X,Y)] = 0$ and $\mathbb{E}_{P_{0,X}}[D_n^U(\widetilde{X})] = 0$. If the influence functions converge in L^2 and n/N converges to a finite limit, the central limit theorem and independence of the two samples yield

$$\sqrt{n}\,(\widehat{\psi} - \psi_0) \rightsquigarrow N\Big(0,\ \lim_{n\to\infty}\Big[\mathrm{Var}_{P_0}\{D_n^L(X,Y)\} + \frac{n}{N}\mathrm{Var}_{P_{0,X}}\{D_n^U(\widetilde{X})\}\Big]\Big).$$

The influence pair thus determines the asymptotic variance, accounting for variability in both samples. The following result characterizes the influence pair for the AIPW class, identifies the efficient choice of q, and derives the semiparametric efficiency bound in the unrestricted two-sample model [8, 75, 74].

Proposition 3 (Influence pair and efficiency of AIPW)**.** *Under Assumption 1, for any fixed square-integrable function $q : \mathcal{X} \to \mathbb{R}$, the estimator $\widehat{\psi}(q)$ is asymptotically linear with triangular influence pair $(D_{q,n}^L, D_{q,n}^U)$, where*

$$D_{q,n}^L(X,Y) = Y - \psi_0 - (1-\rho_n)\{q^\circ(X) - \psi_0\},$$
$$D_{q,n}^U(\widetilde{X}) = (1-\rho_n)\{q^\circ(\widetilde{X}) - \psi_0\},$$

and $q^\circ(X) := q(X) - \mathbb{E}_{P_{0,X}}[q(X)] + \psi_0$ is the recentered prediction function. The asymptotic variance of $\sqrt{n}\{\widehat{\psi}(q) - \psi_0\}$ is

$$\rho_0\mathrm{Var}_{P_0}[\mu_0(X)] + \mathbb{E}_{P_0}\big[\mathrm{Var}[Y \mid X]\big] + (1-\rho_0)\mathrm{Var}_{P_0}[q(X) - \mu_0(X)].$$

This variance is minimized when $q = \mu_0$ up to an additive constant. Consequently, the efficient triangular influence pair in the unrestricted two-sample model is $(D_{\mu_0,n}^L, D_{\mu_0,n}^U)$, with efficiency bound $\rho_0\mathrm{Var}_{P_0}[\mu_0(X)] + \mathbb{E}_{P_0}\big[\mathrm{Var}[Y \mid X]\big]$.

By Proposition 3, the limiting variance of the AIPW estimator depends on the score function q only through the mean squared error of its recentered version q°, namely $\mathbb{E}_{P_0}\big[(q^\circ(X) - \mu_0(X))^2\big]$. Thus, within the AIPW class, efficiency is maximized by choosing q so that q° approximates $\mu_0(X)$ as well as possible, with equality when $q^\circ = \mu_0$, in particular at $q = \mu_0$. This shows that full efficiency generally requires using the full covariate information X, or at least a summary rich enough to

[2] In more general missing-at-random settings with unknown, covariate-dependent missingness probabilities, the analogous AIPW estimator remains doubly robust asymptotically, but it no longer enjoys the same finite-sample unbiasedness as in the present known-ρ_n setting [91].

recover $\mu_0(X)$ [74]. In practice, this motivates taking q to be an estimated prediction score $m(X)$, learned using flexible black-box methods.

AIPW, PPI, and empirical-efficiency maximization. The most common choice is $q(X) = m(X)$, where $m(X)$ is a prediction score, yielding the usual augmented inverse-probability weighted (AIPW) estimator [75]:

$$\widehat{\psi}_{\mathrm{AIPW}} = \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}.$$

Equivalently,

$$\widehat{\psi}_{\mathrm{AIPW}} = \frac{1}{M}\left\{\sum_{i=1}^{n} m(X_i) + \sum_{j=1}^{N} m(\widetilde{X}_j) + \frac{1}{\rho_n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}\right\},$$

where ρ_n^{-1} is the inverse probability of observing the outcome.

Another estimator in this class was considered by Angelopoulos et al. [1] under the name PPI,[3] and we adopt that terminology here. It corresponds to the score choice

$$q(X) = \frac{m(X)}{1-\rho_n}.$$

Substituting this choice into (10) gives

$$\widehat{\psi}_{\mathrm{PPI}} = \mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}.$$

Relative to $\widehat{\psi}_{\mathrm{AIPW}}$, $\widehat{\psi}_{\mathrm{PPI}}$ uses only the unlabeled sample in the plug-in term and rescales the score by the factor $(1-\rho_n)^{-1} > 1$. By Proposition 3, this helps only if the rescaled score is closer to $\mu_0(X)$ in mean squared error. Thus, PPI is inefficient when $m = \mu_0$, and is typically less efficient when m is already reasonably accurate. When ρ_n is small and m is sufficiently inaccurate, however, neither estimator is necessarily more efficient than the other [34, 80].

Because both $\widehat{\psi}_{\mathrm{AIPW}}$ and $\widehat{\psi}_{\mathrm{PPI}}$ use scores in the scaling class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$, one can choose the scaling factor data-adaptively rather than fixing it in advance. One way to do so is to minimize an empirical estimate of the asymptotic variance of $\sqrt{n}\{\widehat{\psi}(q) - \psi_0\}$:

$$\sigma_n^2(q) := \widehat{\mathrm{Var}}_L\{Y - (1-\rho_n)q(X)\} + \rho_n(1-\rho_n)\,\widehat{\mathrm{Var}}_U\{q(\widetilde{X})\},$$

where $\widehat{\mathrm{Var}}_L(a) := \mathbb{P}_n^L[(a - \mathbb{P}_n^L a)^2]$ and $\widehat{\mathrm{Var}}_U(b) := \mathbb{P}_N^U[(b - \mathbb{P}_N^U b)^2]$. That is, one chooses λ to minimize $\sigma_n^2(q)$ over $q(X) = \lambda m(X)$, a procedure known as empirical-efficiency maximization [83, 103]. The resulting estimator, called PPI++ by Angelopoulos et al. [2], is simply the AIPW estimator with an empirically rescaled prediction rule $\hat{\lambda} m(X)$, where $\hat{\lambda}$ estimates the population-optimal scaling coefficient

$$\lambda^\star = \frac{\mathrm{Cov}\big(Y, m(X)\big)}{\mathrm{Var}\big(m(X)\big)}. \tag{11}$$

The same principle extends beyond simple rescaling. For example, one can minimize $\sigma_n^2(q)$ over richer classes such as $\{\lambda m(X) + \beta_{\mathrm{lin}}^\top X : \lambda \in \mathbb{R},\ \beta_{\mathrm{lin}} \in \mathbb{R}^d\}$, over monotone transformations of the score, or over a reproducing kernel Hilbert space (Appendix D). More generally, the same criterion can be used in cross-validation to choose among candidate estimator classes [83].

Regression adjustment and TMLE can yield the efficient influence pair by updating the regression so that its empirical residual mean vanishes [87, 105, 78]. The calibration identity (4) gives an exact instance of this construction.

D Additional remarks

Remark 1 (Implementation)**.** Isotonic regression is widely available in standard software. A standard algorithm is the pooled-adjacent-violators algorithm (PAVA) [6]. Although it is generally fast, it

[3] The PPI and PPI++ estimators are algebraically equivalent to the model-assisted survey-design estimators of Cassel et al. [10] and Särndal et al. [86], respectively; see Mozer [64] for a review.

may become slow at ultra-large sample sizes. A scalable alternative is to use regression tree software with monotonicity constraints, such as XGBoost or LightGBM (see Appendix O). An advantage of tree-based software is that the fit can be regularized by constraining or tuning the number of leaves, the maximum tree depth, or the minimum number of observations per leaf.

Remark 2 (Platt scaling adjustment)**.** The linear-calibration idea extends beyond squared-error loss. One can instead use logistic loss, Poisson loss, or, more generally, losses from canonical-link generalized linear models [78]. In particular, when $Y \in \{0,1\}$ and $m(X) \in (0,1)$ is an initial predictor, one may apply Platt scaling [21, 70] by fitting a logistic regression of Y on the logit of $m(X)$. Writing $\eta_i(a,b) := a\,\mathrm{logit}(m(X_i)) + b$, this amounts to solving

$$\begin{aligned}\big(\hat{a}_{\mathrm{platt}}, \hat{b}_{\mathrm{platt}}\big) \in \arg\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} \Big[-Y_i \log \sigma\big(\eta_i(a,b)\big) \\ -(1-Y_i)\log\Big(1-\sigma\big(\eta_i(a,b)\big)\Big)\Big].\end{aligned} \tag{12}$$

with calibrated predictor

$$q_n^{\mathrm{platt}}(X) := \sigma\big(\hat{a}_{\mathrm{platt}}\,\mathrm{logit}(m(X)) + \hat{b}_{\mathrm{platt}}\big),$$

where $\sigma(t) := (1+e^{-t})^{-1}$. Platt scaling may be preferred over linear regression calibration when predicted probabilities are systematically over- or under-confident, in which case the calibration curve is often approximately sigmoidal. More generally, outcomes can be rescaled to $[0,1]$ before applying this procedure. In online settings, calibration properties of Platt scaling are developed in [31].

D.1 Cross-fitting and leave-one-out calibration

Fitting a calibrator to pooled out-of-fold predictions still uses each labeled outcome in the calibration fit. To hold out both steps, partition the labeled indices into folds $I_1, \ldots, I_K$ independently of the observations, and fit each calibrated score $q^{(-k)}$ without the pairs indexed by I_k. Any predictor fitting and tuning used to construct $q^{(-k)}$ also exclude I_k. Let $k(i)$ be the fold containing i and $c_n := N/(n+N)$. The cross-fitted AIPW estimator is

$$\widehat{\psi}_{\mathrm{CF}} := \frac{1}{n}\sum_{i=1}^{n}\left[Y_i + c_n\{\mathbb{P}_N^U q^{(-k(i))} - q^{(-k(i))}(X_i)\}\right]. \tag{13}$$

For each i, conditional on the training sample, $\mathbb{E}[q^{(-k(i))}(X_i)] = \mathbb{E}[\mathbb{P}_N^U q^{(-k(i))}]$. Thus (13) is unbiased when these predictions are integrable and the fitted scores do not use the unlabeled observations.

Leave-one-out calibration takes $K = n$. Writing q_{-i} for the score fitted without (X_i, Y_i), we obtain

$$\widehat{\psi}_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i + c_n\{\mathbb{P}_N^U q_{-i} - q_{-i}(X_i)\}\right]. \tag{14}$$

The residual correction is retained because the held-out residuals need not average to zero. In the simulation, we use

$$\widehat{\mathrm{se}}^2_{\mathrm{LOO}} = \frac{\widehat{\mathrm{Var}}_L\{Y_i - c_n q_{-i}(X_i)\}}{n} + \frac{\widehat{\mathrm{Var}}_U\{c_n \bar{q}(\widetilde{X}_j)\}}{N}, \qquad \bar{q} := \frac{1}{n}\sum_{i=1}^{n} q_{-i}, \tag{15}$$

where the empirical variances use denominators $n-1$ and $N-1$. The bias, coverage, and interval widths of this procedure are reported in Appendix N.1.

Remark 3 (Empirical efficiency maximization over a prediction class)**.** The empirical variance criterion $\sigma_n^2(q)$ in Section C.1 can be minimized over any candidate class $\mathcal{G} \subset L^2(P_{0,X})$, rather than only over rescalings of one score:

$$\hat{q} \in \arg\min_{q\in\mathcal{G}} \sigma_n^2(q), \qquad \widehat{\psi}_{\mathrm{eem}} := \widehat{\psi}(\hat{q}).$$

At the $\sqrt{n}$ scale, Proposition 3 gives the limiting variance

$$V(q) = \rho_0 \mathrm{Var}[Y] + (1-\rho_0)\mathrm{Var}[Y - q(X)].$$

Writing $q^\circ = q - \mathbb{E}_{P_{0,X}} q + \psi_0$, we have $\mathrm{Var}[Y - q(X)] = \mathbb{E}_{P_0}[(Y - q^\circ(X))^2]$. If $\mathcal{G}$ is closed under constant shifts, population least-squares regression therefore targets a variance-minimizing score. For tuning, average held-out estimates of the influence-function variance across folds and use the selected class or tuning parameter in a cross-fitted AIPW estimator [83].

Remark 4 (Shrinkage via Venn–Abers)**.** When the labeled sample is small, isotonic calibration may be unstable. A natural alternative is Venn–Abers calibration, originally introduced for binary classification [107]; see [95, 94] for generalized versions beyond the binary setting. Assume $Y \in [0, 1]$, or rescale otherwise. For each x, augment the labeled calibration sample with the hypothetical point $(m(x), y)$ for $y \in \{0, 1\}$, fit the resulting isotonic calibrators $f_n^{(x,0)}$ and $f_n^{(x,1)}$, and form the interval-valued prediction

$$\big[f_n^{(x,0)}\{m(x)\},\ f_n^{(x,1)}\{m(x)\}\big].$$

For the Venn–Abers comparator, we shrink the interval midpoint toward the unadjusted AIPW estimator:

$$q_n^{\mathrm{VA}}(x) := q_n^{\mathrm{mid}}(x) + \big\{f_n^{(x,1)}\{m(x)\} - f_n^{(x,0)}\{m(x)\}\big\}\big\{\widehat{\psi}_{AIPW} - q_n^{\mathrm{mid}}(x)\big\},$$

where

$$q_n^{\mathrm{mid}}(x) := \frac{f_n^{(x,1)}\{m(x)\} + f_n^{(x,0)}\{m(x)\}}{2}.$$

The shrinkage weight is the width of the Venn–Abers interval.

Remark 5 (Intercept-only calibration)**.** Set $\hat{a}_{\mathrm{const}} := \mathbb{P}_n^L\{Y - m(X)\}$ and $q_n^{\mathrm{const}} := m + \hat{a}_{\mathrm{const}}$. Then PPI is the unlabeled mean $\mathbb{P}_N^U q_n^{\mathrm{const}}$, whereas AIPW is the pooled mean $\mathbb{P}_M^X q_n^{\mathrm{const}}$.

E Calibrated DML for randomized trials

Let $A \in \{0, 1\}$ denote randomized treatment and $Y = AY(1) + (1 - A)Y(0)$. Assume consistency, no interference, and $A \perp (X, Y(1), Y(0))$, with positive assignment probabilities. When estimating $\mu_a := \mathbb{E}[Y(a)]$, outcomes from arm a are labeled, and covariates from the other arm form the unlabeled sample. Let $\mathbb{P}_{n,a}$ denote the empirical mean in arm a, $\mathbb{P}_n$ the pooled mean, and $\hat{\pi}_a = n_a/n$. If the calibrated arm-specific score satisfies $\mathbb{P}_{n,a}\{Y - m_{n,a}^\star(X)\} = 0$, then

$$\hat{\mu}_a^{\mathrm{cal}} := \mathbb{P}_n\{m_{n,a}^\star(X)\} \tag{16}$$

$$= \mathbb{P}_n\{m_{n,a}^\star(X)\} + \mathbb{P}_n\left[\frac{\mathbf{1}(A = a)}{\hat{\pi}_a}\{Y - m_{n,a}^\star(X)\}\right]. \tag{17}$$

The residual correction is zero because it equals $\mathbb{P}_{n,a}\{Y - m_{n,a}^\star(X)\}$. Thus (4) applies to each arm, and the calibrated treatment-effect estimator is

$$\hat{\tau}^{\mathrm{cal}} := \hat{\mu}_1^{\mathrm{cal}} - \hat{\mu}_0^{\mathrm{cal}}. \tag{18}$$

These arm-specific updates are instances of calibrated DML [100].

F Moment equations under missing at random

For a target satisfying $\mathbb{E}[\phi_{\theta_0}(X, Y)] = 0$, let R indicate an observed outcome, $\pi_0(X) = \Pr(R = 1 \mid X) > 0$, and $a_\theta(X) = \mathbb{E}[\phi_\theta(X, Y) \mid X]$. Under $R \perp Y \mid X$, the augmented moment equation is

$$\mathbb{P}_M\left[a_\theta(X) + \frac{R}{\pi_0(X)}\{\phi_\theta(X, Y) - a_\theta(X)\}\right] = 0.$$

If calibration makes the empirical residual term vanish, this equation reduces to $\mathbb{P}_M a_\theta = 0$. Extensions with estimated missingness probabilities are studied in Robins et al. [75, 76], van der Laan et al. [100].

G Pooled i.i.d. Missing-Data Formulation

When $\rho_0 > 0$, the joint two-sample setup can be embedded into a pooled i.i.d. model by introducing a sample-membership indicator $R \in \{0, 1\}$, where $R = 1$ denotes labeled observations and $R = 0$ denotes unlabeled observations. In that formulation one observes i.i.d. data $O = (X, R, RY)$ with $\mathbb{P}(R = 1) = \rho_0$, and the calibrated estimator becomes

$$\widehat{\psi}_{\text{cal}} = \mathbb{P}_M(q_n) + \mathbb{P}_M\left[\frac{R}{\rho_0}\{Y - q_n\}\right]$$

up to the same calibration identity as in the main text. The efficient influence function in the reduced model observing only the score is

$$\mathbb{E}[Y \mid m(X)] - \psi_0 + \frac{R}{\rho_0}\{Y - \mathbb{E}[Y \mid m(X)]\},$$

which is the single-sample analogue of the efficient influence pair obtained by applying proposition 3 to the reduced model with scalar covariate $m(X)$. Thus the usual i.i.d. missing-data formulation is a convenient special case, but the main text keeps the two-dataset structure explicit.

H AIPW comparison identity

Lemma 1 (AIPW comparison identity)**.** *Suppose* $\mathbb{P}_n^L\{Y - q_n(X)\} = 0$*. For any comparison score* g *with finite empirical averages,*

$$\widehat{\psi}_{\text{cal}} = \widehat{\psi}(g) + (1 - \rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)(q_n - g).$$

Proof. The residual equation gives $\widehat{\psi}_{\text{cal}} = \widehat{\psi}(q_n)$. Subtracting $\widehat{\psi}(g)$ and collecting terms proves the identity. □

I PPI++ and linear calibration

Proof of Proposition 2. Write

$$\hat{v}_L := \mathbb{P}_n^L[(m - \mathbb{P}_n^L m)^2], \quad \hat{v}_U := \mathbb{P}_N^U[(m - \mathbb{P}_N^U m)^2], \quad \hat{c}_L := \mathbb{P}_n^L[(Y - \mathbb{P}_n^L Y)(m - \mathbb{P}_n^L m)].$$

On the event $\hat{v}_L > 0$, the least-squares normal equations give

$$\hat{a}_{\text{lin}} = \frac{\hat{c}_L}{\hat{v}_L}, \qquad \hat{b}_{\text{lin}} = \mathbb{P}_n^L Y - \hat{a}_{\text{lin}}\mathbb{P}_n^L m.$$

Differentiating the empirical variance criterion in Section C.1 gives

$$\hat{\lambda}_{++} = \frac{\hat{c}_L}{(1 - \rho_n)\hat{v}_L + \rho_n \hat{v}_U}.$$

Finite second moments and $\text{Var}[m(X)] > 0$ imply $\hat{v}_L, \hat{v}_U \to_p \text{Var}[m(X)]$ and $\hat{c}_L \to_p \text{Cov}(Y, m(X))$. Thus $\Pr(\hat{v}_L > 0) \to 1$, both coefficients converge to $a_0 := \text{Cov}(Y, m(X))/\text{Var}[m(X)]$, and $\hat{\lambda}_{++} - \hat{a}_{\text{lin}} = o_p(1)$.

For $\Delta_n := \mathbb{P}_N^U m - \mathbb{P}_n^L m$, substituting the fitted intercept gives

$$\widehat{\psi}_{\text{lin}} = \mathbb{P}_n^L Y + (1 - \rho_n)\hat{a}_{\text{lin}}\Delta_n, \qquad \widehat{\psi}_{++} = \mathbb{P}_n^L Y + (1 - \rho_n)\hat{\lambda}_{++}\Delta_n.$$

Subtracting proves the exact identity in the proposition. Independence and finite second moments give $\Delta_n = O_p(n^{-1/2} + N^{-1/2})$. Since $\rho_n \to \rho_0 < 1$, the ratio n/N is bounded, so

$$\widehat{\psi}_{++} - \widehat{\psi}_{\text{lin}} = (1 - \rho_n)o_p(1)\Delta_n = o_p(n^{-1/2}).$$

□

J Proof of Proposition 3

Proof of Proposition 3. Adding a constant to q leaves $\widehat{\psi}(q)$ unchanged. Thus $\widehat{\psi}(q) = \widehat{\psi}(q^\circ)$, where $q^\circ = q - \mathbb{E}_{P_{0,X}} q + \psi_0$. Centering the empirical means gives the exact identity

$$\widehat{\psi}(q) - \psi_0 = (\mathbb{P}_n^L - P_0)D_{q,n}^L + (\mathbb{P}_N^U - P_{0,X})D_{q,n}^U.$$

Both components have mean zero. Write $A = \mu_0 - \psi_0$, $\delta = q^\circ - \mu_0$, and $\varepsilon = Y - \mu_0$. Then

$$D_{q,n}^L = \varepsilon + \rho_n A - (1-\rho_n)\delta, \qquad D_{q,n}^U = (1-\rho_n)(A+\delta).$$

Since $\mathbb{E}[\varepsilon \mid X] = 0$ and $n/N = \rho_n/(1-\rho_n)$, the covariance terms cancel in

$$\mathrm{Var}(D_{q,n}^L) + \frac{n}{N}\mathrm{Var}(D_{q,n}^U) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \rho_n \mathrm{Var}(\mu_0) + (1-\rho_n)\mathrm{Var}(\delta).$$

Taking limits gives the stated variance, since $\mathrm{Var}(\delta) = \mathrm{Var}(q - \mu_0)$. The choice $q = \mu_0$ removes the last term and gives the missing-data efficiency bound [76, 74, 93]. □

K Supporting lemmas for calibration

Lemma 2 (Blockwise residual orthogonality)**.** *Let $B_1, \ldots, B_J$ be the pooled adjacent violator blocks of the isotonic fit, and let $q_n^{\mathrm{iso}}(X_i) = c_j$ for $i \in B_j$. Then*

$$\sum_{i \in B_j} (Y_i - c_j) = 0, \qquad j = 1, \ldots, J.$$

Consequently, for every measurable function $h : \mathbb{R} \to \mathbb{R}$ such that $h(q_n^{\mathrm{iso}}(X_i))$ is finite for all i,

$$\sum_{i=1}^n h(q_n^{\mathrm{iso}}(X_i))\{Y_i - q_n^{\mathrm{iso}}(X_i)\} = 0.$$

Proof. The fitted values from isotonic regression are constant on pooled adjacent violator blocks. The KKT conditions imply that residuals sum to zero on each block. Since $h(q_n^{\mathrm{iso}}(X_i)) = h(c_j)$ for all $i \in B_j$,

$$\sum_{i=1}^n h(q_n^{\mathrm{iso}}(X_i))\{Y_i - q_n^{\mathrm{iso}}(X_i)\} = \sum_{j=1}^J h(c_j) \sum_{i \in B_j} (Y_i - c_j) = 0.$$

□

Proof of proposition 1. Lemma 2 gives (5). By definition, the fitted isotonic score minimizes the empirical squared loss among all nondecreasing transformations of m, proving the risk comparison. Taking the identity transformation gives the comparison with the raw score. □

Lemma 3 (Bounded variation after isotonic post-processing)**.** *Suppose Assumption 3 holds. Let $T := m(X)$ and $S_n := q_n^{\mathrm{iso}}(X)$. Define*

$$\eta(t) := \mathbb{E}_{P_0}[Y \mid T = t], \qquad \eta_n(s) := \mathbb{E}_{P_0}[Y \mid S_n = s].$$

If $t \mapsto \eta(t)$ has total variation at most V, then, conditional on the labeled sample, the map $s \mapsto \eta_n(s)$ has total variation at most V. If, in addition, $|\eta(t)| \le C_0$ on the support of T, then $|\eta_n(s)| \le C_0$ on the support of S_n.

Proof. We adapt the proof of Lemma 6 in [98]. Let $S_n = f_n(T)$, where f_n is the fitted isotonic map. Conditional on the labeled sample, f_n is nondecreasing. Hence, for each s in the support of S_n, there exists an interval $B_s \subset \mathbb{R}$, possibly a singleton, such that

$$\{S_n = s\} = \{T \in B_s\}.$$

Therefore

$$\eta_n(s) = \mathbb{E}_{P_0}[Y \mid S_n = s] = \mathbb{E}_{P_0}[\eta(T) \mid T \in B_s].$$

Let $\eta = \eta^+ - \eta^-$ be the Jordan decomposition of η, where η^+ and η^- are nondecreasing and

$$\|\eta\|_{TV} = \|\eta^+\|_{TV} + \|\eta^-\|_{TV} \le V.$$

Then

$$\eta_n(s) = \mathbb{E}_{P_0}[\eta^+(T) \mid T \in B_s] - \mathbb{E}_{P_0}[\eta^-(T) \mid T \in B_s].$$

If $s_1 < s_2$, then monotonicity of f_n implies every point of B_{s_1} is no larger than every point of B_{s_2}. Hence the conditional distribution of $T \mid T \in B_{s_2}$ first-order stochastically dominates that of $T \mid T \in B_{s_1}$. Therefore, for any nondecreasing function g,

$$\mathbb{E}_{P_0}[g(T) \mid T \in B_{s_1}] \le \mathbb{E}_{P_0}[g(T) \mid T \in B_{s_2}].$$

Applying this with $g = \eta^+$ and $g = \eta^-$, both maps

$$s \mapsto \mathbb{E}_{P_0}[\eta^+(T) \mid T \in B_s] \qquad \text{and} \qquad s \mapsto \mathbb{E}_{P_0}[\eta^-(T) \mid T \in B_s]$$

are nondecreasing.

Moreover, each of these conditional expectations takes values in the range of the corresponding function, so

$$\big\|s \mapsto \mathbb{E}_{P_0}[\eta^+(T) \mid T \in B_s]\big\|_{TV} \le \|\eta^+\|_{TV}, \qquad \big\|s \mapsto \mathbb{E}_{P_0}[\eta^-(T) \mid T \in B_s]\big\|_{TV} \le \|\eta^-\|_{TV}.$$

Hence

$$\|\eta_n\|_{TV} \le \|\eta^+\|_{TV} + \|\eta^-\|_{TV} = \|\eta\|_{TV} \le V.$$

Finally, if $|\eta(t)| \le C_0$ on the support of T, then

$$|\eta_n(s)| = |\mathbb{E}_{P_0}[\eta(T) \mid T \in B_s]| \le C_0$$

for all s in the support of S_n. □

Lemma 4 (The population isotonic projection is centered)**.** *Under Assumptions 1 and 2,* $\mathbb{E}_{P_{0,X}}[q_0^{\mathrm{iso}}(X)] = \psi_0$.

Proof. The isotonic class is closed under constant shifts, so $c \mapsto \mathbb{E}[(Y - q_0^{\mathrm{iso}}(X) - c)^2]$ is minimized at $c = 0$. Its derivative there is $-2\mathbb{E}[Y - q_0^{\mathrm{iso}}(X)]$, giving the stated identity. □

Proof convention. In the following proofs, fix $\zeta > 0$. By $\sup_x |q_n^{\mathrm{iso}}(x)| = O_p(1)$, there exist a deterministic constant $C = C_\zeta \ge C_0$ and $n_0 < \infty$ such that the event

$$A_{n,C} := \Big\{\sup_x |q_n^{\mathrm{iso}}(x)| \le C\Big\}$$

satisfies $P_0(A_{n,C}) \ge 1 - \zeta$ for all $n \ge n_0$. We carry out the deterministic bounded-class arguments on $A_{n,C}$, where all implicit constants may depend on C but not on n. Since ζ is arbitrary, the resulting rates and distributional statements hold unconditionally.

Lemma 5 (Isotonic L^2 rate)**.** *Suppose Assumptions 1 and 2 holds. Then*

$$\|q_n^{\mathrm{iso}} - q_0^{\mathrm{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Proof of Lemma 5. Let $T := m(X)$, $\eta(T) := \mathbb{E}_{P_0}[Y \mid T]$, and

$$\mathcal{M}_C := \{f(T) : f \in \mathcal{F}_{\mathrm{iso}},\ \|f(T)\|_\infty \le C\}.$$

By Assumption 2, $\eta(T)$ is bounded by $C_0 \le C$, and on $A_{n,C}$ we have $q_n^{\mathrm{iso}} \in \mathcal{M}_C$. Since clipping any monotone candidate to $[-C, C]$ preserves monotonicity and can only decrease squared risk against $\eta(T)$, the population isotonic projection q_0^{iso} also belongs to $\mathcal{M}_C$. For every $m \in \mathcal{M}_C$,

$$\mathbb{E}_{P_0}[(Y - m)^2] = \mathbb{E}_{P_0}[(Y - \eta)^2] + \mathbb{E}_{P_0}[(\eta - m)^2],$$

which shows that q_0^{iso} is the $L^2(P_T)$-projection of η onto the closed convex set $\mathcal{M}_C$.

Now define the centered class

$$\mathcal{F}_0 := \{g = m - q_0^{\mathrm{iso}} :\ m \in \mathcal{M}_C\}.$$

For $g \in \mathcal{F}_0$, let $\ell_g(Z) := (Y - q_0^{\mathrm{iso}}(X) - g(X))^2 - (Y - q_0^{\mathrm{iso}}(X))^2$, where $Z = (X, Y)$. Since q_n^{iso} minimizes the empirical squared risk over $\mathcal{F}_{\mathrm{iso}}$, on $A_{n,C}$ the random element $\hat{g} := q_n^{\mathrm{iso}} - q_0^{\mathrm{iso}}$ belongs to $\mathcal{F}_0$ and satisfies

$$\mathbb{P}_n^L\{\ell_{\hat{g}}\} \le 0 = \mathbb{P}_n^L\{\ell_0\}.$$

Moreover,

$$\mathbb{E}_{P_0}[\ell_g] = \|g\|_{2,P_{0,X}}^2 - 2\mathbb{E}_{P_0}[(Y - q_0^{\mathrm{iso}})g].$$

Because g is measurable with respect to T, the tower property gives

$$\mathbb{E}_{P_0}[(Y - q_0^{\mathrm{iso}})g] = \mathbb{E}_{P_0}[(\eta - q_0^{\mathrm{iso}})g].$$

Since q_0^{iso} is the $L^2(P_T)$-projection of η onto the convex set $\mathcal{M}_C$, the Hilbert-space projection inequality yields

$$\mathbb{E}_{P_0}[(\eta - q_0^{\mathrm{iso}})(m - q_0^{\mathrm{iso}})] \le 0 \qquad \text{for every } m \in \mathcal{M}_C.$$

Hence

$$\mathbb{E}_{P_0}[\ell_g] \ge \|g\|_{2,P_{0,X}}^2,$$

so squared loss has the required quadratic margin around q_0^{iso}.

We next bound the modulus of continuity of the empirical process indexed by the localized classes

$$\mathcal{F}_\delta := \{g \in \mathcal{F}_0 : \|g\|_{2,P_{0,X}} \le \delta\}.$$

The bracketing bound for bounded monotone functions gives

$$\log N_{[]}(\varepsilon, \mathcal{F}_\delta, L_2(P_{0,X})) \lesssim \varepsilon^{-1}, \qquad 0 < \varepsilon < \delta.$$

Consequently,

$$J_{[]}(\delta, \mathcal{F}_\delta, L_2(P_{0,X})) \lesssim \delta^{1/2};$$

see, e.g., van der Vaart and Wellner [106].

Write $\xi := Y - q_0^{\mathrm{iso}}(X)$, so that $\ell_g = g^2 - 2\xi g$. By Condition (ii) and the bounds on η, q_0^{iso}, and g, there exists $A_C < \infty$ such that

$$\mathbb{E}_{P_0}[|\xi g|^k] \le k! A_C^k \|g\|_{2,P_{0,X}}^2, \qquad g \in \mathcal{F}_0, \quad k \ge 2. \tag{19}$$

The same inequality holds for bounded bracket widths. Thus multiplication by ξ maps $L^2(P_{0,X})$ brackets to Bernstein-norm brackets with widths bounded by a constant multiple of their original widths; see Lemmas 8–9 of Bibaut and van der Laan [7]. The squared class $\{g^2 : g \in \mathcal{F}_\delta\}$ is a bounded Lipschitz image of $\mathcal{F}_\delta$. Applying Lemmas 3.4.2–3.4.3 of van der Vaart and Wellner [106] to these two classes gives

$$E^\star \sup_{g \in \mathcal{F}_\delta} |(\mathbb{P}_n^L - P_0)\ell_g| \lesssim n^{-1/2}\delta^{1/2} + n^{-1}\delta^{-1}. \tag{20}$$

For $\delta \ge n^{-1/3}$, the right-hand side is bounded by a constant times $n^{-1/2}\delta^{1/2}$.

Set $\varepsilon_n := n^{-1/3}$. The empirical risk inequality, the quadratic margin, and (20) imply, by peeling into shells $2^s\varepsilon_n \le \|\hat{g}\|_{2,P_{0,X}} < 2^{s+1}\varepsilon_n$ and applying Markov's inequality, that

$$\begin{aligned}\Pr\{\|\hat{g}\|_{2,P_{0,X}} \ge 2^S\varepsilon_n,\ A_{n,C}\} &\lesssim \sum_{s \ge S} \frac{n^{-1/2}(2^{s+1}\varepsilon_n)^{1/2}}{2^{2s}\varepsilon_n^2} \\ &\lesssim \sum_{s \ge S} 2^{-3s/2} \longrightarrow 0 \qquad (S \to \infty).\end{aligned}$$

Hence

$$\|q_n^{\mathrm{iso}} - q_0^{\mathrm{iso}}\|_{2,P_{0,X}} = O_p(n^{-1/3}), \qquad \|q_n^{\mathrm{iso}} - q_0^{\mathrm{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Since $P_0(A_{n,C}^c) \le \zeta$ for all large n, and $\zeta > 0$ was arbitrary, the same rate holds unconditionally, proving Lemma 5. □

Theorem 3 (Calibration error of isotonic post-processing)**.** *Suppose Assumptions 1 to 3 hold. Let*

$$q_{0|n}(X) := \mathbb{E}_{P_0}[Y \mid q_n^{\mathrm{iso}}(X)].$$

Then

$$\|q_{0|n} - q_n^{\mathrm{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Proof. The proof follows that of Theorem 1 in [98]; see also [94]. Let

$$T := m(X), \qquad S_n := q_n^{\mathrm{iso}}(X), \qquad \eta_n(s) := \mathbb{E}_{P_0}[Y \mid S_n = s],$$

and define

$$\Delta_n(X) := q_{0|n}(X) - q_n^{\mathrm{iso}}(X) = \eta_n(S_n) - S_n.$$

Conditional on the labeled sample, $h_n(s) := \eta_n(s) - s$ is a measurable function of s. Applying proposition 1 pathwise with $h = h_n$ yields

$$\mathbb{P}_n^L\{\Delta_n(X)\{Y - S_n\}\} = 0.$$

Also,

$$\mathbb{E}_{P_0}[\Delta_n(X)\{Y - S_n\}] = \mathbb{E}_{P_0}[\Delta_n(X)\,\mathbb{E}_{P_0}[Y - S_n \mid S_n]] = \mathbb{E}_{P_0}[\Delta_n^2] = \|\Delta_n\|_{2,P_{0,X}}^2.$$

Hence

$$\|\Delta_n\|_{2,P_{0,X}}^2 = (\mathbb{E}_{P_0} - \mathbb{P}_n^L)\{\Delta_n(X)\{Y - S_n\}\}.$$

Fix $\zeta > 0$, and work on the event $A_{n,C}$ from the proof convention. On this event, S_n is uniformly bounded by C. By Lemma 3, the function η_n belongs to the bounded-variation class

$$\mathcal{F}_{\mathrm{TV},C} := \{f : \mathbb{R} \to \mathbb{R} : \|f\|_\infty \le C,\ \|f\|_{TV} \le V\},$$

where $V < \infty$ is the variation bound in Assumption 3. Let

$$\mathcal{F}_{\mathrm{iso},C} := \{f : \mathbb{R} \to \mathbb{R} : f \text{ nondecreasing and } \|f\|_\infty \le C\}.$$

Since $S_n = f_n(T)$ for a nondecreasing fitted isotonic map f_n, and composition with a monotone map preserves bounded variation, the function

$$t \mapsto \eta_n\{f_n(t)\}$$

also belongs to $\mathcal{F}_{\mathrm{TV},C}$, after enlarging C if necessary. Thus $\eta_n(S_n) = h_1(X)$ for some $h_1 = f_1 \circ m$ with $f_1 \in \mathcal{F}_{\mathrm{TV},C}$, while $S_n = h_2(X)$ for $h_2 = f_2 \circ m$ with $f_2 = f_n \in \mathcal{F}_{\mathrm{iso},C}$.

Define the localized product class

$$\mathcal{G}_{n,C,\delta} := \Big\{(h_1 - h_2)(Y - h_2) : h_1 = f_1 \circ m,\ h_2 = f_2 \circ m,\\ f_1 \in \mathcal{F}_{\mathrm{TV},C},\ f_2 \in \mathcal{F}_{\mathrm{iso},C},\ \|h_1 - h_2\|_{2,P_{0,X}} \le \delta\Big\}.$$

Then

$$\Delta_n(X)\{Y - S_n\} \in \mathcal{G}_{n,C,\delta_n}, \qquad \delta_n := \|\Delta_n\|_{2,P_{0,X}}.$$

Write $\eta(T) := \mathbb{E}_{P_0}[Y \mid T]$ and $\varepsilon := Y - \eta(T)$. Each element of $\mathcal{G}_{n,C,\delta}$ decomposes as

$$(h_1 - h_2)(Y - h_2) = (h_1 - h_2)\varepsilon + (h_1 - h_2)\{\eta(T) - h_2\}.$$

Bounded-variation functions are differences of bounded monotone functions, so their bracketing entropy is $O(\epsilon^{-1})$. The second term is a bounded Lipschitz transformation of (h_1, h_2). For the first term, Condition (ii) gives the moment bound in (19), with $h_1 - h_2$ in place of g and ε in place of ξ. Its Bernstein-norm bracket widths are therefore controlled by $L^2(P_{0,X})$ bracket widths. The second term has $L^2(P_0)$ norm $O(\delta)$, and the first has Bernstein norm $O(\delta)$; both bracketing entropy integrals are $O(\delta^{1/2})$. Lemmas 3.4.2–3.4.3 of van der Vaart and Wellner [106] give

$$E^\star \sup_{g \in \mathcal{G}_{n,C,\delta}} |(\mathbb{P}_n^L - P_0)\{g\}| \lesssim n^{-1/2}\delta^{1/2} + n^{-1}\delta^{-1}.$$

For $\delta \ge n^{-1/3}$, this is bounded by a constant times $n^{-1/2}\delta^{1/2}$.

Set $\varepsilon_n := n^{-1/3}$. For $s \ge 0$, let

$$B_s := \{2^s\varepsilon_n \le \delta_n < 2^{s+1}\varepsilon_n\}.$$

On $B_s \cap A_{n,C}$,

$$2^{2s}\varepsilon_n^2 \le \delta_n^2 \le \sup_{g \in \mathcal{G}_{n,C,2^{s+1}\varepsilon_n}} |(\mathbb{P}_n^L - P_0)\{g\}|.$$

Therefore, by Markov's inequality,

$$P_0(B_s \cap A_{n,C}) \lesssim \frac{E^\star \sup_{g \in \mathcal{G}_{n,C,2^{s+1}\varepsilon_n}} |(\mathbb{P}_n^L - P_0)\{g\}|}{2^{2s}\varepsilon_n^2} \lesssim \frac{n^{-1/2}(2^{s+1}\varepsilon_n)^{1/2}}{2^{2s}\varepsilon_n^2} \lesssim 2^{-3s/2},$$

since $n^{-1/2}\varepsilon_n^{-3/2} = 1$. Summing over $s \geq S$ yields

$$P_0(\delta_n \geq 2^S \varepsilon_n,\ A_{n,C}) \lesssim \sum_{s=S}^{\infty} 2^{-3s/2} \xrightarrow[S\to\infty]{} 0.$$

Because $P_0(A_{n,C}^c) \leq \zeta$ for all large n, and $\zeta > 0$ is arbitrary, it follows that

$$\|\Delta_n\|_{2,P_{0,X}} = O_p(n^{-1/3}), \qquad \|\Delta_n\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

This is exactly the claimed calibration-error rate. □

Lemma 6 (Centered empirical-process bounds for isotonic calibration)**.** *Suppose Assumptions 1 and 2 holds. Then*

$$(\mathbb{P}_n^L - P_{0,X})\{q_n^{\text{iso}}(X) - q_0^{\text{iso}}(X)\} = O_p(n^{-2/3}),$$
$$(\mathbb{P}_N^U - P_{0,X})\{q_n^{\text{iso}}(\widetilde{X}) - q_0^{\text{iso}}(\widetilde{X})\} = O_p\big(n^{-1/3}N^{-1/2}\big).$$

Proof of Lemma 6. By Lemma 5,

$$\|q_n^{\text{iso}} - q_0^{\text{iso}}\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

To control the centered empirical-process term, work on the event $A_{n,C}$ from the proof convention. For fixed $R < \infty$, consider the localized class

$$\mathcal{F}_{n,C} := \{g = f \circ m - q_0^{\text{iso}} :\ f \in \mathcal{F}_{\text{iso}},\ \|f \circ m\|_\infty \leq C,\ \|g\|_{2,P_{0,X}} \leq Rn^{-1/3}\}.$$

By Lemma 5, the probability of $A_{n,C} \cap \{q_n^{\text{iso}} - q_0^{\text{iso}} \notin \mathcal{F}_{n,C}\}$ can be made arbitrarily small by taking R large. Because m is fixed, $\mathcal{F}_{n,C}$ is a translated class of bounded monotone functions of a one-dimensional score. The monotone bracketing bound implies

$$J_{[]}(r, \mathcal{F}_{n,C}, L_2(P_0)) \lesssim r^{1/2}.$$

Applying Lemma 3.4.2 of van der Vaart and Wellner [106] at radius $r_n = Rn^{-1/3}$ yields

$$\sup_{g \in \mathcal{F}_{n,C}} |(\mathbb{P}_n^L - P_{0,X})\{g\}| = O_p\Big(n^{-1/2}r_n^{1/2} + n^{-1}r_n^{-1}\Big) = O_p(n^{-2/3}).$$

In particular,

$$(\mathbb{P}_n^L - P_{0,X})\{q_n^{\text{iso}} - q_0^{\text{iso}}\} = O_p(n^{-2/3}).$$

Since $P_0(A_{n,C}^c) \leq \zeta$ for all large n, and $\zeta > 0$ is arbitrary, the same bound holds unconditionally.

For the unlabeled sample, conditional on the labeled data, $q_n^{\text{iso}} - q_0^{\text{iso}}$ is fixed and the unlabeled sample is independent, so

$$(\mathbb{P}_N^U - P_{0,X})\{q_n^{\text{iso}} - q_0^{\text{iso}}\} = O_p\left(N^{-1/2}\|q_n^{\text{iso}} - q_0^{\text{iso}}\|_{2,P_{0,X}}\right) = O_p\big(n^{-1/3}N^{-1/2}\big)$$

by Lemma 5. □

Lemma 7 (Centered empirical-process bounds for isotonic post-processing)**.** *Suppose Assumptions 1 to 3 hold. Let*

$$\Delta_n(X) := q_{0|n}(X) - q_n^{\text{iso}}(X).$$

Then

$$\big(\mathbb{P}_n^L - P_{0,X}\big)\{\Delta_n(X)\} = O_p(n^{-2/3}), \qquad \big(\mathbb{P}_N^U - P_{0,X}\big)\{\Delta_n(\widetilde{X})\} = O_p\big(n^{-1/3}N^{-1/2}\big).$$

Proof. By theorem 3,

$$\|\Delta_n\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Let $V < \infty$ be the variation bound in Assumption 3. Fix $\zeta > 0$, and work on the event $A_{n,C}$. On this event, q_n^{iso} is uniformly bounded by C. By Lemma 3, the function

$$s \mapsto \eta_n(s) := \mathbb{E}_{P_0}[Y \mid q_n^{\text{iso}}(X) = s]$$

has total variation at most V and is bounded by C on the support of $q_n^{\text{iso}}(X)$. Since $q_n^{\text{iso}}(X) = f_n(T)$ with $T = m(X)$ and f_n nondecreasing, it follows that

$$q_{0|n}(X) = \eta_n\{f_n(T)\}$$

is a bounded-variation transform of T, while $q_n^{\text{iso}}(X)$ is a bounded monotone transform of T.

Consider the classes

$$\mathcal{F}_{\text{TV},C} := \{f : \mathbb{R} \to \mathbb{R} : \|f\|_\infty \le C,\ \|f\|_{TV} \le V\},$$
$$\mathcal{F}_{\text{iso},C} := \{f : \mathbb{R} \to \mathbb{R} : f \text{ nondecreasing and } \|f\|_\infty \le C\}.$$

For fixed $R < \infty$, define

$$\mathcal{H}_{n,C} := \Big\{h_1 - h_2 : h_1 = f_1 \circ m,\ h_2 = f_2 \circ m,\ f_1 \in \mathcal{F}_{\text{TV},C},\ f_2 \in \mathcal{F}_{\text{iso},C},\ \|h_1 - h_2\|_{2,P_{0,X}} \le Rn^{-1/3}\Big\}.$$

By theorem 3, the probability of $A_{n,C} \cap \{\Delta_n \notin \mathcal{H}_{n,C}\}$ can be made arbitrarily small by taking R large.

Because bounded-variation functions are differences of bounded monotone functions, the localized bracketing entropy of $\mathcal{H}_{n,C}$ obeys the same bound as in the proof of theorem 3, namely

$$J_{[]}(\delta, \mathcal{H}_{n,C}, L_2(P_0)) \lesssim \delta^{1/2}.$$

Applying Lemma 3.4.2 of van der Vaart and Wellner [106] at radius $r_n = Rn^{-1/3}$ gives

$$\sup_{h \in \mathcal{H}_{n,C}} |(\mathbb{P}_n^L - P_{0,X})h| = O_p\Big(n^{-1/2}r_n^{1/2} + n^{-1}r_n^{-1}\Big) = O_p(n^{-2/3}).$$

Taking R large gives the same bound for $(\mathbb{P}_n^L - P_{0,X})\Delta_n$ on $A_{n,C}$. Since $P_0(A_{n,C}^c) \le \zeta$ for all large n, and $\zeta > 0$ was arbitrary, the same bound holds unconditionally.

For the unlabeled sample, conditional on the labeled data, Δ_n is fixed and the unlabeled sample is independent. Hence

$$\big(\mathbb{P}_N^U - P_{0,X}\big)\{\Delta_n(\widetilde{X})\} = O_p\left(N^{-1/2}\|\Delta_n\|_{2,P_{0,X}}\right) = O_p\big(n^{-1/3}N^{-1/2}\big)$$

by theorem 3. □

L Proof of asymptotic linearity

Proof of theorem 1. By Lemma 4, $\mathbb{E}_{P_{0,X}}[q_0^{\text{iso}}] = \psi_0$. The isotonic fit has mean-zero labeled residuals by Lemma 2. Applying Lemma 1 with $g = q_0^{\text{iso}}$ gives

$$\widehat{\psi}_{\text{iso}} = \rho_n\mathbb{P}_n^L\left\{q_0^{\text{iso}} + \rho_n^{-1}(Y - q_0^{\text{iso}})\right\} + (1-\rho_n)\mathbb{P}_N^U\{q_0^{\text{iso}}(\widetilde{X})\} + (1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{q_n^{\text{iso}} - q_0^{\text{iso}}\}.$$

Subtracting ψ_0 gives

$$\begin{aligned}\widehat{\psi}_{\text{iso}} - \psi_0 = (\mathbb{P}_n^L - P_0)D^L_{q_0^{\text{iso}},n} + (\mathbb{P}_N^U - P_{0,X})D^U_{q_0^{\text{iso}},n} \\ + (1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{q_n^{\text{iso}} - q_0^{\text{iso}}\}.\end{aligned}$$

Also,

$$\begin{aligned}(1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{q_n^{\text{iso}} - q_0^{\text{iso}}\} = (1-\rho_n)(\mathbb{P}_N^U - P_{0,X})\{q_n^{\text{iso}} - q_0^{\text{iso}}\} \\ - (1-\rho_n)(\mathbb{P}_n^L - P_{0,X})\{q_n^{\text{iso}} - q_0^{\text{iso}}\}.\end{aligned}$$

By Lemma 6, the labeled term is $O_p(n^{-2/3})$ and the unlabeled term is $O_p(n^{-1/3}N^{-1/2})$. Hence the remainder satisfies the bound in equation (9). Moreover,

$$\sqrt{n}(1-\rho_n)n^{-2/3} \le n^{-1/6}$$

and, writing $t_n = N/n$,

$$\sqrt{n}(1-\rho_n)n^{-1/3}N^{-1/2} = n^{-1/3}\frac{\sqrt{t_n}}{1+t_n} \le \frac{1}{2}n^{-1/3}.$$

Therefore $R_{n,N} = o_p(n^{-1/2})$, which proves equation (9). □

Proof of corollary 1. By theorem 1,

$$\sqrt{n}\,(\widehat{\psi}_{\text{iso}} - \psi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D^L_{q_0^{\text{iso}},n}(X_i, Y_i) + \frac{\sqrt{n}}{N}\sum_{j=1}^{N} D^U_{q_0^{\text{iso}},n}(\widetilde{X}_j) + o_p(1).$$

The labeled and unlabeled samples are independent and each summand has mean zero. Moreover,

$$\frac{n}{N}\text{Var}\{D^U_{q_0^{\text{iso}},n}(\widetilde{X})\} = \rho_n(1-\rho_n)\text{Var}\{q_0^{\text{iso}}(X)\}.$$

Since $\rho_n \to \rho_0 \in [0,1)$, a triangular-array central limit theorem yields the stated asymptotic normal law with variance σ_0^2.

For the variance estimator, let $g_n := q_n^{\text{iso}} - q_0^{\text{iso}}$. By Lemma 5,

$$\mathbb{E}_{P_{0,X}}[g_n^2] = \|g_n\|^2_{2,P_{0,X}} = O_p(n^{-2/3}).$$

Applying Lemma 3.4.2 of van der Vaart and Wellner [106] to the squared class gives

$$\sup_{g\in\mathcal{F}_\delta} |(\mathbb{P}_n^L - P_{0,X})\{g^2\}| = O_p\Big(n^{-1/2}\delta^{1/2} + n^{-1}\delta^{-1}\Big),$$

where $\mathcal{F}_\delta$ is the localized class from the proof of Lemma 5. By that lemma, on $A_{n,C}$ the function g_n belongs to $\mathcal{F}_{Rn^{-1/3}}$ with arbitrarily high probability for sufficiently large fixed R. Taking $\delta = Rn^{-1/3}$ gives

$$(\mathbb{P}_n^L - P_{0,X})\{g_n^2\} = O_p(n^{-2/3}).$$

Hence

$$\mathbb{P}_n^L(g_n^2) = O_p(n^{-2/3}) = o_p(1).$$

On the boundedness event $A_{n,C}$, conditional on the labeled data, g_n is fixed and $|g_n| \le 2C$, so

$$(\mathbb{P}_N^U - P_{0,X})\{g_n^2\} = O_p\left(N^{-1/2}\|g_n^2\|_{2,P_{0,X}}\right) = O_p\left(C\,N^{-1/2}\|g_n\|_{2,P_{0,X}}\right) = O_p\big(n^{-1/3}N^{-1/2}\big).$$

Therefore

$$\mathbb{P}_N^U(g_n^2) = O_p(n^{-2/3}) + O_p\big(n^{-1/3}N^{-1/2}\big) = o_p(1).$$

Since $P_0(A_{n,C}^c) \le \zeta$ for all large n and arbitrary $\zeta > 0$, these bounds hold unconditionally. Also, $\widehat{\psi}_{\text{iso}} \to_p \psi_0$ follows from theorem 1.

Finally,

$$\widehat{D}^L - D^L_{q_0^{\text{iso}},n} = -(1-\rho_n)g_n - \rho_n(\widehat{\psi}_{\text{iso}} - \psi_0),$$

and

$$\widehat{D}^U - D^U_{q_0^{\text{iso}},n} = (1-\rho_n)\{g_n(\widetilde{X}) - (\widehat{\psi}_{\text{iso}} - \psi_0)\}.$$

Because $\mathbb{P}_n^L(g_n^2), \mathbb{P}_N^U(g_n^2) = o_p(1)$ and $\widehat{\psi}_{\text{iso}} \to_p \psi_0$, it follows that

$$\mathbb{P}_n^L\Big[\{\widehat{D}^L - D^L_{q_0^{\text{iso}},n}\}^2\Big] = o_p(1), \qquad \frac{n}{N}\mathbb{P}_N^U\Big[\{\widehat{D}^U - D^U_{q_0^{\text{iso}},n}\}^2\Big] = o_p(1),$$

where the second conclusion uses

$$\frac{n}{N}(1-\rho_n)^2 = \rho_n(1-\rho_n) \le \frac{1}{4}.$$

The law of large numbers and $\rho_n \to \rho_0$ also give

$$\mathbb{P}_n^L\{(D^L_{q_0^{\text{iso}},n})^2\} + \frac{n}{N}\mathbb{P}_N^U\{(D^U_{q_0^{\text{iso}},n})^2\} \;\to_p\; \sigma_0^2.$$

The empirical means of both estimated influence components converge to zero, so within-sample centering does not change this limit. Therefore $\widehat{\sigma}^2 \to_p \sigma_0^2$. The Wald interval claim then follows by Slutsky's theorem. □

M Proof of post-processing equivalence and reduced-model efficiency

Proof of Theorem 2. Set $\Delta_n = q_{0|n} - q_n^{\mathrm{iso}}$. Lemma 1 with $g = q_{0|n}$ gives

$$\widehat{\psi}_{\mathrm{eff}} - \widehat{\psi}_{\mathrm{iso}} = (1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\Delta_n.$$

By Lemma 7, the right-hand side is

$$O_p((1-\rho_n)n^{-2/3}) + O_p((1-\rho_n)n^{-1/3}N^{-1/2}) = o_p(n^{-1/2}),$$

using the same bounds as in the proof of Theorem 1. The calibration-error rate is Theorem 3. Together these prove both conclusions. □

Proof of corollary 2. For part (i), if $q_0^{\mathrm{iso}} = \mu_0$ almost surely, then the asymptotic influence pair in theorem 1 is

$$D^L_{\mu_0,n}(X,Y) = Y - \psi_0 - (1-\rho_n)\{\mu_0(X) - \psi_0\}, \qquad D^U_{\mu_0,n}(\widetilde{X}) = (1-\rho_n)\{\mu_0(\widetilde{X}) - \psi_0\}.$$

By proposition 3, this is the efficient influence pair in the unrestricted two-sample model. Therefore $\widehat{\psi}_{\mathrm{iso}}$ is efficient for ψ_0 in the full two-sample experiment.

For part (ii), fix a square-integrable nondecreasing transformation $g = \theta \circ m$, and let $g^\circ = g - \mathbb{E}_{P_{0,X}} g + \psi_0$. Adding a constant preserves monotonicity, so population isotonic optimality and Lemma 4 give

$$\mathrm{Var}(Y - q_0^{\mathrm{iso}}) = \mathbb{E}[(Y - q_0^{\mathrm{iso}})^2] \le \mathbb{E}[(Y - g^\circ)^2] = \mathrm{Var}(Y - g).$$

Proposition 3 now gives the efficiency comparison.

For part (iii), let

$$S := m(X), \qquad \eta_0(s) := \mathbb{E}_{P_0}[Y \mid S = s].$$

proposition 3 applied to the reduced two-sample experiment with scalar covariate S shows that the efficient triangular influence pair in that reduced experiment is

$$D^L_{\eta_0,n}(S,Y) = Y - \psi_0 - (1-\rho_n)\{\eta_0(S) - \psi_0\}, \qquad D^U_{\eta_0,n}(\widetilde{S}) = (1-\rho_n)\{\eta_0(\widetilde{S}) - \psi_0\}.$$

If η_0 is monotone increasing, then η_0 belongs to the isotonic class and

$$\mathbb{E}_{P_0}\left[\{Y - g(S)\}^2\right] = \mathbb{E}_{P_0}\left[\{Y - \eta_0(S)\}^2\right] + \mathbb{E}_{P_0}\left[\{\eta_0(S) - g(S)\}^2\right]$$

for every measurable g. Hence the population isotonic regression of Y on the score $S = m(X)$ is exactly $\eta_0(S)$, so

$$q_0^{\mathrm{iso}}(X) = \eta_0\{m(X)\} \qquad \text{almost surely.}$$

Therefore, the influence pair in theorem 1 coincides with the reduced-model efficient influence pair above, so $\widehat{\psi}_{\mathrm{iso}}$ is efficient for ψ_0 in the reduced experiment indexed by $m(X)$. The isotonic class is closed under constant shifts and positive rescaling. The corresponding population first-order conditions give $\mathbb{E}[Y - q_0^{\mathrm{iso}}] = \mathbb{E}[q_0^{\mathrm{iso}}(Y - q_0^{\mathrm{iso}})] = 0$, so $\mathrm{Cov}(Y, q_0^{\mathrm{iso}}) = \mathrm{Var}(q_0^{\mathrm{iso}})$. Substitution into σ_0^2 gives the stated variance reduction relative to the labeled mean. □

N Additional empirical details

N.1 Small-sample comparisons and RePPI

The small-sample experiment uses $S \sim N(0,1)$, $p(S) = \{1 + \exp(-5S)\}^{-1}$, and $Y \mid S \sim \mathrm{Bernoulli}\{p(S)\}$. Let $C(t) = \min\{0.99, \max(0.01, t)\}$, $\sigma(t) = (1+e^{-t})^{-1}$, and $Z \sim N(0,1)$ independently of (S,Y). The mildly miscalibrated score is $C[\sigma\{1.5\,\mathrm{logit}(C[p(S) + 0.005Z])\}]$; the strongly miscalibrated score is $C[-0.15 + 0.75\,\sigma\{3\,\mathrm{logit}(p(S)) - 0.5\}]$. The simulation seed is 20260411. Figure 3 reports signed bias, coverage of nominal 95% intervals, interval width, and relative efficiency versus PPI++.

We implement two scalar-score versions of RePPI Algorithm 1 [45], using three-way sample splitting and weights proportional to the evaluation-fold sizes. RePPI-RF uses a 100-tree random forest. RePPI-MLP-CV selects hidden-layer widths from (8), $(32,32)$, and $(32,32,32)$, and penalties from

$\{10^{-2}, 10^{-4}\}$, by two-fold validation MSE within each outer training fold. Both estimate the scalar power-tuning coefficient on the evaluation fold, use the global factor $N/(n+N)$, and set the coefficient to zero when fitted-score variance is numerically singular. The neural implementation uses the feature vector $[1, m(X)]$. Its prediction map and the scalar single-fold calculation were checked against the public implementation; the three-way aggregation follows the paper's Algorithm 1.

RePPI-RF has lower RMSE than PPI++ in one of 14 settings; RePPI-MLP-CV does so in five. For $n \geq 50$, the neural estimator's RMSE is 0.576–0.668 times that of the labeled mean. At $n = 20$, the ratios are 1.450 and 9.690 under mild and strong miscalibration. Neither RePPI implementation has lower RMSE than smooth monotone calibration in these settings. The largest neural-estimator error occurs when a fitted-score variance of 1.04×10^{-7} produces a coefficient of 1238.2. All replications are retained. Singular-variance fallbacks occur in 175 random-forest replications and 46 neural-network replications; the neural fits produce 2,120 convergence warnings.

This appendix describes the benchmark construction used in section 5. We compare PPI with AIPW, clipped PPI++ with AIPW-EM, and `LinearCal` with `AutoCal`.

We use the official `ppi_py` real-data mean-estimation examples and preserve their labeled-sample-size grids. The `forest` experiment uses $n \in \{50, 100, \ldots, 500\}$. The `galaxies` and semisupervised `census_income` benchmarks use $n \in \{50, 155, 261, 366, 472, 577, 683, 788, 894, 1000\}$. For each n, the remaining observations form the unlabeled sample. We rerandomize this split 500 times at each n, always evaluating all estimators on the same split and taking the full-sample mean as the benchmark truth.

We compare the labeled-only mean, PPI, AIPW, PPI++, AIPW-EM, `LinearCal`, `MonoSpline`, `IsoCal`, and `AutoCal`. The binary-outcome benchmarks also include imputation, Platt scaling, and Venn–Abers shrinkage. PPI and PPI++ use the official `ppi_py` implementations. PPI++ restricts its coefficient to $[0, 1]$, whereas AIPW-EM minimizes the empirical variance without that constraint.

Calibration maps are fitted on the labeled pairs. We use the calibrated scores in the AIPW estimator, retaining the residual correction after clipping or binning. `LinearCal` fits an affine map; `MonoSpline` fits a smooth nondecreasing spline with a roughness penalty. Both clip predictions to the range of the labeled outcomes. `IsoCal` uses a minimum bin size of 10. Platt scaling fits logistic regression to a stabilized logit transform of the score. `AutoCal` selects among AIPW, linear calibration, monotone splines, and isotonic calibration using 20-fold cross-validation of the estimated influence-function variance. Each fold uses an unlabeled subsample of size $\min(N, 10n)$ to evaluate this criterion. We report Monte Carlo bias, variance, MSE, Wald-interval coverage, and relative efficiency versus PPI.

N.2 LLM score ordering and held-out calibration

For MATH responses evaluated by ArmoRM, empirical correctness increases across twelve score bins, from 0.078 in the lowest bin to 0.920 in the highest. Platt calibration fitted on 100 randomly selected labels reduces mean squared prediction error from 0.4683 to 0.2020 on the remaining held-out observations, a reduction of 56.9%.

N.3 LLM evaluator breakdown

The largest raw-score MSE inflation occurs for Skywork and Athene. The other evaluator summaries show smaller differences among the main-text estimators. The evaluator-specific results show where calibration provides the largest gains.

N.4 LLM ranking supplement

The PPE human-preference track also induces a downstream ranking problem across the selected target models. Using the same repeated labeled/unlabeled splits as in section 5, we therefore summarize how well each estimator recovers the target-model ordering induced by the full-data human win shares. For each evaluator, labeled-sample size, replicate, and estimator, we rank the eight selected target models by estimated win share, compare that ordering against the ranking induced by the full-data targets, and then macro-average across evaluators. We report Spearman rank correlation, top-1 identification rate, and top-1 regret, defined as the difference in true win share between the best

Algorithm 1 Isotonic-calibrated PPI

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$
2: Fit $\hat{f} \in \arg\min_{f \in \mathcal{F}_{\text{iso}}} \sum_{i=1}^n \{Y_i - f(T_i)\}^2$
3: Set $q_n^{\text{iso}}(X_i) = \hat{f}(T_i)$ and $q_n^{\text{iso}}(\widetilde{X}_j) = \hat{f}(\widetilde{T}_j)$
4: Form $\widehat{\psi}_{\text{iso}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n q_n^{\text{iso}}(X_i) + \sum_{j=1}^N q_n^{\text{iso}}(\widetilde{X}_j)\}$
5: Set $\widehat{D}_i^L = Y_i - \widehat{\psi}_{\text{iso}} - (1-\rho_n)\{q_n^{\text{iso}}(X_i) - \widehat{\psi}_{\text{iso}}\}$ and $\widehat{D}_j^U = (1-\rho_n)\{q_n^{\text{iso}}(\widetilde{X}_j) - \widehat{\psi}_{\text{iso}}\}$
6: Set $\overline{D}_L = n^{-1}\sum_{i=1}^n \widehat{D}_i^L$ and $\overline{D}_U = N^{-1}\sum_{j=1}^N \widehat{D}_j^U$
7: Set $\widehat{\text{SE}}_{\text{iso}}^2 = n^{-2}\sum_{i=1}^n (\widehat{D}_i^L - \overline{D}_L)^2 + N^{-2}\sum_{j=1}^N (\widehat{D}_j^U - \overline{D}_U)^2$
8: Return $\widehat{\psi}_{\text{iso}} \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\text{iso}}$

target model and the model selected by the estimator. The main pattern mirrors the mean-estimation results: calibrated and efficiency-maximized estimators substantially outperform raw PPI/AIPW, while remaining close to labeled-only at small n. At $n = 200$, `AutoCal` attains the highest average Spearman correlation, about 0.905; at $n = 400$, `LinearCal` gives the best overall ranking quality, with Spearman about 0.941 and top-1 identification about 0.646.

N.5 Data provenance, compute, and reproducibility

All empirical results use public benchmark sources cited in the main text and appendix. The tabular reproductions use the official `ppi_py` mean-estimation datasets; the LLM experiments use PPE Human Preference V1 and PPE correctness data derived from MMLU-Pro, MATH, GPQA, IFEval, and MBPP+. The submission does not introduce a new dataset or foundation model. Figures and tables are generated from the experiment scripts and cached CSV summaries; raw public data and evaluator-score sources remain governed by their upstream licenses and terms.

The main simulation uses 500 Monte Carlo repetitions per setting, and the small-sample comparison uses 1,000. The reproduced PPI benchmarks use 500 random labeled/unlabeled splits for each combination of dataset and labeled sample size. The PPE experiments use 200 random splits for each combination of evaluator, benchmark, and target model. We use public precomputed evaluator scores.

The local runs used Python 3.9.6 with `numpy` 2.0.2, `pandas` 2.3.3, `scikit-learn` 1.6.1, `xgboost` 2.1.4, `lightgbm` 4.6.0, `matplotlib` 3.9.4, and `ppi-python` 0.2.3. Computations used an Apple M1 Pro machine with 10 logical CPUs and 16 GB RAM. The 1,000-repetition small-sample experiment used eight parallel workers and took 9,214 seconds of elapsed time.

O Implementation sketch

Reference implementations are in `experiments/` in the supplementary code. The algorithms below use the notation of the paper.

O.1 Algorithmic summaries

Algorithm 2 Linearly calibrated PPI

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$
2: Fit $(\hat{a}, \hat{b}) \in \arg\min_{a,b} \sum_{i=1}^n \{Y_i - (aT_i + b)\}^2$
3: Set $q_n^{\text{lin}}(X_i) = \hat{a}T_i + \hat{b}$ and $q_n^{\text{lin}}(\widetilde{X}_j) = \hat{a}\widetilde{T}_j + \hat{b}$
4: Form $\widehat{\psi}_{\text{lin}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n q_n^{\text{lin}}(X_i) + \sum_{j=1}^N q_n^{\text{lin}}(\widetilde{X}_j)\}$
5: Set $\widehat{D}_i^L = Y_i - \widehat{\psi}_{\text{lin}} - (1-\rho_n)\{q_n^{\text{lin}}(X_i) - \widehat{\psi}_{\text{lin}}\}$ and $\widehat{D}_j^U = (1-\rho_n)\{q_n^{\text{lin}}(\widetilde{X}_j) - \widehat{\psi}_{\text{lin}}\}$
6: Set $\overline{D}_L = n^{-1}\sum_{i=1}^n \widehat{D}_i^L$ and $\overline{D}_U = N^{-1}\sum_{j=1}^N \widehat{D}_j^U$
7: Set $\widehat{\text{SE}}_{\text{lin}}^2 = n^{-2}\sum_{i=1}^n (\widehat{D}_i^L - \overline{D}_L)^2 + N^{-2}\sum_{j=1}^N (\widehat{D}_j^U - \overline{D}_U)^2$
8: Return $\widehat{\psi}_{\text{lin}} \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\text{lin}}$

Algorithm 3 Linearly calibrated PPI with covariate adjustment

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$
2: Fit $(\hat{a}, \hat{\beta}, \hat{b}) \in \arg\min_{a,\beta,b} \sum_{i=1}^n \{Y_i - (aT_i + X_i^\top \beta + b)\}^2$
3: Set $q_n^{\text{lin}}(X_i) = \hat{a}T_i + X_i^\top\hat{\beta} + \hat{b}$ and $q_n^{\text{lin}}(\widetilde{X}_j) = \hat{a}\widetilde{T}_j + \widetilde{X}_j^\top\hat{\beta} + \hat{b}$
4: Form $\widehat{\psi}_{\text{lin}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n q_n^{\text{lin}}(X_i) + \sum_{j=1}^N q_n^{\text{lin}}(\widetilde{X}_j)\}$
5: Set $\widehat{D}_i^L = Y_i - \widehat{\psi}_{\text{lin}} - (1-\rho_n)\{q_n^{\text{lin}}(X_i) - \widehat{\psi}_{\text{lin}}\}$ and $\widehat{D}_j^U = (1-\rho_n)\{q_n^{\text{lin}}(\widetilde{X}_j) - \widehat{\psi}_{\text{lin}}\}$
6: Set $\overline{D}_L = n^{-1}\sum_{i=1}^n \widehat{D}_i^L$ and $\overline{D}_U = N^{-1}\sum_{j=1}^N \widehat{D}_j^U$
7: Set $\widehat{\text{SE}}_{\text{lin}}^2 = n^{-2}\sum_{i=1}^n (\widehat{D}_i^L - \overline{D}_L)^2 + N^{-2}\sum_{j=1}^N (\widehat{D}_j^U - \overline{D}_U)^2$
8: Return $\widehat{\psi}_{\text{lin}} \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\text{lin}}$

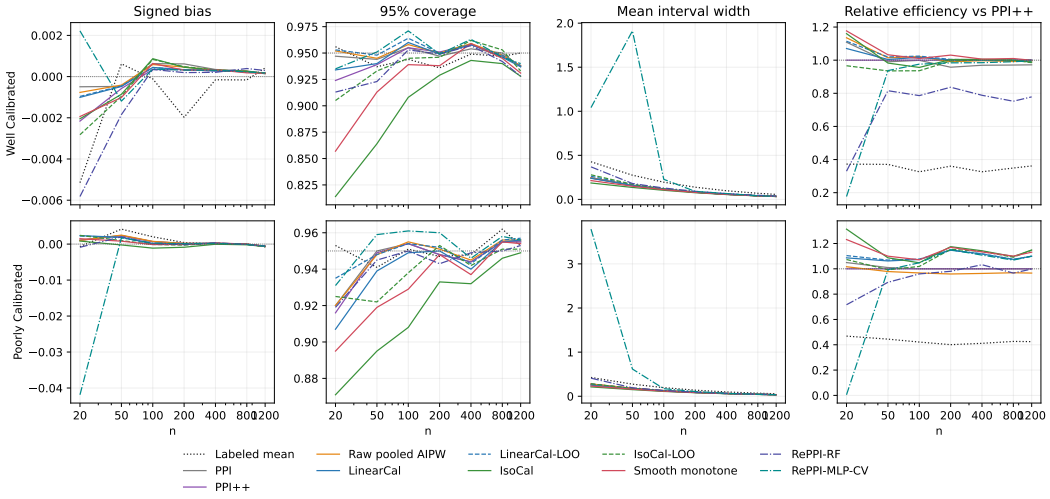

Figure 3: Signed bias, coverage of nominal 95% intervals, interval width, and relative efficiency versus PPI++ in the small-sample experiment. The numerical summary includes Monte Carlo standard errors.

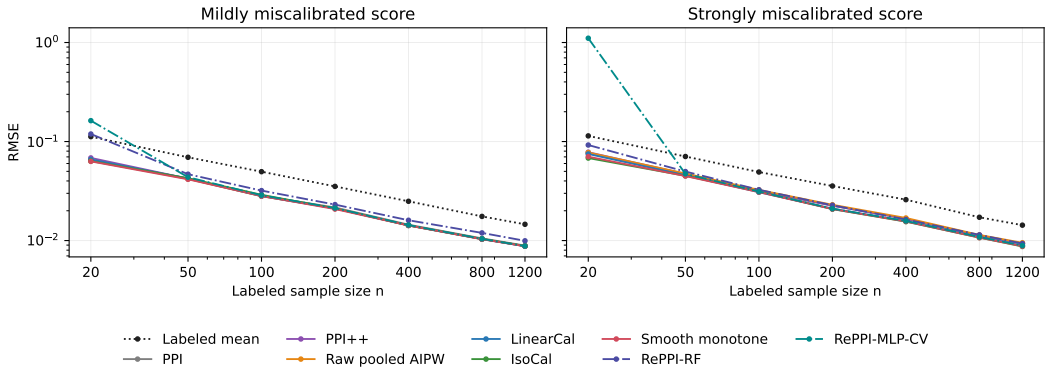

Figure 4: Small-sample RMSE comparison including the random-forest and cross-validated neural-network RePPI implementations.

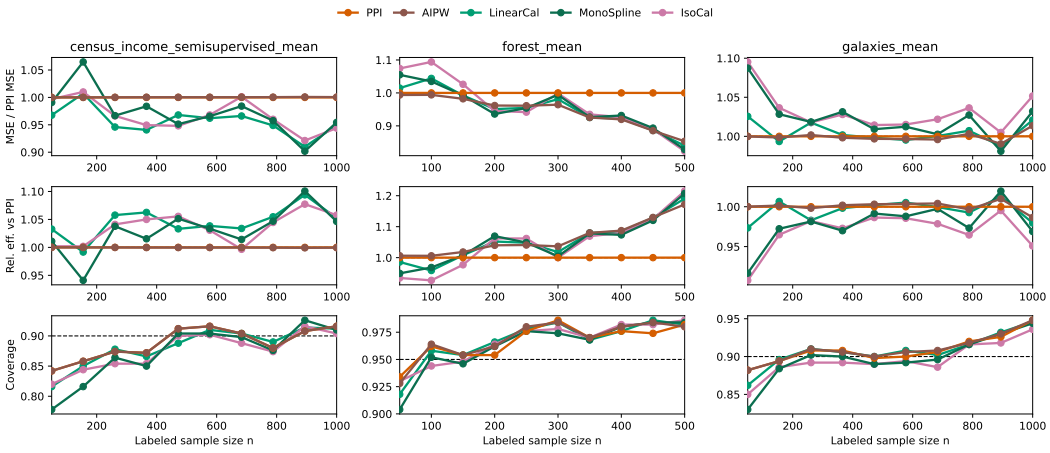

Figure 5: Calibration-focused appendix comparison across the reproduced PPI benchmarks, restricted to PPI, AIPW, `LinearCal`, `MonoSpline`, and `IsoCal`. The panels report normalized MSE relative to PPI, relative efficiency versus PPI, and coverage across labeled-sample-size regimes.

Dataset	n	N	Estimator	Bias	Variance	MSE	Coverage	RelEff
census_income_semisupervised_mean	50	380041	AIPW	356.4905	69836665.4410	69963750.9366	0.842	1.000
census_income_semisupervised_mean	50	380041	AIPW-EM	260.9833	67728426.6042	67796538.8913	0.822	1.031
census_income_semisupervised_mean	50	380041	AutoCal	-123.2501	69242944.1124	69258134.7078	0.832	1.009
census_income_semisupervised_mean	50	380041	Labeled-only	685.3532	105558402.2427	106028111.3137	0.860	0.662
census_income_semisupervised_mean	50	380041	IsoCal	-215.3832	69697400.2434	69743790.1536	0.820	1.002
census_income_semisupervised_mean	50	380041	LinearCal	260.9294	67622601.4006	67690685.5430	0.816	1.033
census_income_semisupervised_mean	50	380041	MonoSpline	-482.0477	69082916.1270	69315286.1508	0.778	1.011
census_income_semisupervised_mean	50	380041	PPI	356.4473	69834865.9486	69961920.5977	0.842	1.000
census_income_semisupervised_mean	50	380041	PPI++	150.8664	72953053.8328	72975814.4980	0.830	0.957
census_income_semisupervised_mean	577	379514	AIPW	199.8136	6353884.5833	6393810.0648	0.916	1.000
census_income_semisupervised_mean	577	379514	AIPW-EM	177.6067	6114918.2974	6146462.4536	0.912	1.039
census_income_semisupervised_mean	577	379514	AutoCal	134.3746	6201029.6459	6219086.1828	0.910	1.024
census_income_semisupervised_mean	577	379514	Labeled-only	213.2465	9407595.6062	9453069.6946	0.902	0.675
census_income_semisupervised_mean	577	379514	IsoCal	155.0503	6162964.7883	6187005.3977	0.902	1.031
census_income_semisupervised_mean	577	379514	LinearCal	179.8005	6116832.9552	6149161.1594	0.910	1.038
census_income_semisupervised_mean	577	379514	MonoSpline	162.1952	6146069.6864	6172376.9541	0.904	1.033
census_income_semisupervised_mean	577	379514	PPI	199.7932	6351836.6166	6391753.9369	0.916	1.000
census_income_semisupervised_mean	577	379514	PPI++	198.7409	6362047.6345	6401545.5819	0.916	0.998
census_income_semisupervised_mean	1000	379091	AIPW	92.1091	3462051.4638	3470535.5425	0.916	0.999
census_income_semisupervised_mean	1000	379091	AIPW-EM	61.4705	3288235.4312	3292014.0588	0.912	1.052
census_income_semisupervised_mean	1000	379091	AutoCal	25.8207	3289467.9320	3290134.6381	0.922	1.052
census_income_semisupervised_mean	1000	379091	Labeled-only	165.4763	5293296.2822	5320678.6872	0.914	0.654
census_income_semisupervised_mean	1000	379091	IsoCal	51.2697	3271622.4413	3274251.0223	0.904	1.058
census_income_semisupervised_mean	1000	379091	LinearCal	60.7501	3287796.5296	3291487.1058	0.912	1.052
census_income_semisupervised_mean	1000	379091	MonoSpline	60.9626	3306575.6859	3310292.1289	0.910	1.046
census_income_semisupervised_mean	1000	379091	PPI	91.9155	3460071.7259	3468520.1894	0.916	1.000
census_income_semisupervised_mean	1000	379091	PPI++	91.9155	3460071.7259	3468520.1894	0.916	1.000
forest_mean	50	1546	AIPW	-0.0012	0.0018	0.0018	0.928	1.006
forest_mean	50	1546	AIPW-EM	-0.0018	0.0019	0.0019	0.918	0.969
forest_mean	50	1546	AutoCal	-0.0028	0.0019	0.0019	0.924	0.965
forest_mean	50	1546	Labeled-only	0.0017	0.0024	0.0024	0.954	0.762
forest_mean	50	1546	Imputation	-0.0733	0.0000	0.0054	0.000	–
forest_mean	50	1546	IsoCal	-0.0030	0.0020	0.0020	0.930	0.934
forest_mean	50	1546	LinearCal	-0.0016	0.0019	0.0019	0.918	0.986
forest_mean	50	1546	MonoSpline	-0.0021	0.0019	0.0019	0.904	0.949
forest_mean	50	1546	Platt	-0.0003	0.0019	0.0019	0.900	0.948
forest_mean	50	1546	PPI	-0.0013	0.0018	0.0018	0.934	1.000
forest_mean	50	1546	PPI++	-0.0064	0.0019	0.0019	0.928	0.969
forest_mean	50	1546	Venn-Abers	-0.0027	0.0020	0.0020	0.888	0.923
forest_mean	300	1296	AIPW	0.0011	0.0002	0.0002	0.984	1.036
forest_mean	300	1296	AIPW-EM	0.0007	0.0002	0.0002	0.982	1.012
forest_mean	300	1296	AutoCal	0.0006	0.0002	0.0002	0.978	1.002
forest_mean	300	1296	Labeled-only	0.0007	0.0003	0.0004	0.968	0.706
forest_mean	300	1296	Imputation	-0.0733	0.0000	0.0054	0.000	–
forest_mean	300	1296	IsoCal	0.0010	0.0002	0.0002	0.978	1.000
forest_mean	300	1296	LinearCal	0.0010	0.0002	0.0002	0.984	1.018
forest_mean	300	1296	MonoSpline	0.0010	0.0002	0.0002	0.974	1.004
forest_mean	300	1296	Platt	0.0010	0.0002	0.0002	0.982	1.020
forest_mean	300	1296	PPI	0.0011	0.0002	0.0002	0.986	1.000
forest_mean	300	1296	PPI++	-0.0002	0.0002	0.0002	0.980	0.991
forest_mean	300	1296	Venn-Abers	0.0011	0.0002	0.0002	0.978	1.015
forest_mean	500	1096	AIPW	0.0001	0.0001	0.0001	0.980	1.172
forest_mean	500	1096	AIPW-EM	-0.0002	0.0001	0.0001	0.980	1.188
forest_mean	500	1096	AutoCal	-0.0003	0.0001	0.0001	0.986	1.180
forest_mean	500	1096	Labeled-only	0.0001	0.0002	0.0002	0.972	0.879
forest_mean	500	1096	Imputation	-0.0733	0.0000	0.0054	0.000	–
forest_mean	500	1096	IsoCal	-0.0000	0.0001	0.0001	0.986	1.215
forest_mean	500	1096	LinearCal	0.0001	0.0001	0.0001	0.982	1.191
forest_mean	500	1096	MonoSpline	0.0000	0.0001	0.0001	0.984	1.207
forest_mean	500	1096	Platt	0.0001	0.0001	0.0001	0.984	1.203
forest_mean	500	1096	PPI	0.0000	0.0002	0.0002	0.982	1.000
forest_mean	500	1096	PPI++	-0.0006	0.0001	0.0001	0.978	1.179
forest_mean	500	1096	Venn-Abers	-0.0001	0.0001	0.0001	0.978	1.206
galaxies_mean	50	16693	AIPW	0.0035	0.0018	0.0018	0.882	1.000
galaxies_mean	50	16693	AIPW-EM	0.0035	0.0019	0.0020	0.864	0.944
galaxies_mean	50	16693	AutoCal	0.0019	0.0018	0.0018	0.896	0.997
galaxies_mean	50	16693	Labeled-only	0.0014	0.0041	0.0041	0.890	0.451
galaxies_mean	50	16693	Imputation	-0.0433	0.0000	0.0019	0.000	–
galaxies_mean	50	16693	IsoCal	-0.0011	0.0020	0.0020	0.850	0.908
galaxies_mean	50	16693	LinearCal	0.0030	0.0019	0.0019	0.862	0.973
galaxies_mean	50	16693	MonoSpline	0.0026	0.0020	0.0020	0.830	0.917
galaxies_mean	50	16693	Platt	0.0033	0.0019	0.0019	0.856	0.966
galaxies_mean	50	16693	PPI	0.0035	0.0018	0.0018	0.882	1.000
galaxies_mean	50	16693	PPI++	-0.0020	0.0020	0.0020	0.864	0.914
galaxies_mean	50	16693	Venn-Abers	0.0022	0.0020	0.0020	0.830	0.907
galaxies_mean	577	16166	AIPW	-0.0006	0.0001	0.0001	0.906	1.004
galaxies_mean	577	16166	AIPW-EM	-0.0007	0.0001	0.0001	0.906	1.005
galaxies_mean	577	16166	AutoCal	-0.0008	0.0001	0.0001	0.918	0.989
galaxies_mean	577	16166	Labeled-only	-0.0011	0.0003	0.0003	0.910	0.436
galaxies_mean	577	16166	Imputation	-0.0433	0.0000	0.0019	0.000	–
galaxies_mean	577	16166	IsoCal	-0.0007	0.0001	0.0001	0.894	0.986
galaxies_mean	577	16166	LinearCal	-0.0007	0.0001	0.0001	0.908	1.005
galaxies_mean	577	16166	MonoSpline	-0.0006	0.0001	0.0001	0.892	0.988
galaxies_mean	577	16166	Platt	-0.0007	0.0001	0.0001	0.896	1.002
galaxies_mean	577	16166	PPI	-0.0006	0.0001	0.0001	0.900	1.000
galaxies_mean	577	16166	PPI++	-0.0012	0.0001	0.0001	0.902	0.988
galaxies_mean	577	16166	Venn-Abers	-0.0006	0.0001	0.0001	0.892	0.986
galaxies_mean	1000	15743	AIPW	0.0002	0.0001	0.0001	0.948	0.987
galaxies_mean	1000	15743	AIPW-EM	0.0002	0.0001	0.0001	0.948	0.980
galaxies_mean	1000	15743	AutoCal	0.0001	0.0001	0.0001	0.944	0.977
galaxies_mean	1000	15743	Labeled-only	0.0003	0.0002	0.0002	0.884	0.357
galaxies_mean	1000	15743	Imputation	-0.0433	0.0000	0.0019	0.000	–
galaxies_mean	1000	15743	IsoCal	0.0001	0.0001	0.0001	0.936	0.951
galaxies_mean	1000	15743	LinearCal	0.0002	0.0001	0.0001	0.946	0.980
galaxies_mean	1000	15743	MonoSpline	0.0002	0.0001	0.0001	0.944	0.969
galaxies_mean	1000	15743	Platt	0.0002	0.0001	0.0001	0.942	0.979
galaxies_mean	1000	15743	PPI	0.0002	0.0001	0.0001	0.948	1.000
galaxies_mean	1000	15743	PPI++	-0.0002	0.0001	0.0001	0.946	0.977
galaxies_mean	1000	15743	Venn-Abers	0.0002	0.0001	0.0001	0.942	0.952

Table 2: Representative numerical summary at the smallest, middle, and largest labeled sample sizes for each reproduced benchmark. The reported quantities are Monte Carlo bias, empirical variance of the point estimator, MSE, Wald-interval coverage, and relative efficiency versus PPI for the displayed benchmark set, including PPI++, AIPW-EM, `AutoCal`, `MonoSpline`, and `IsoCal`. Relative efficiency is not reported when the estimator has zero variance.

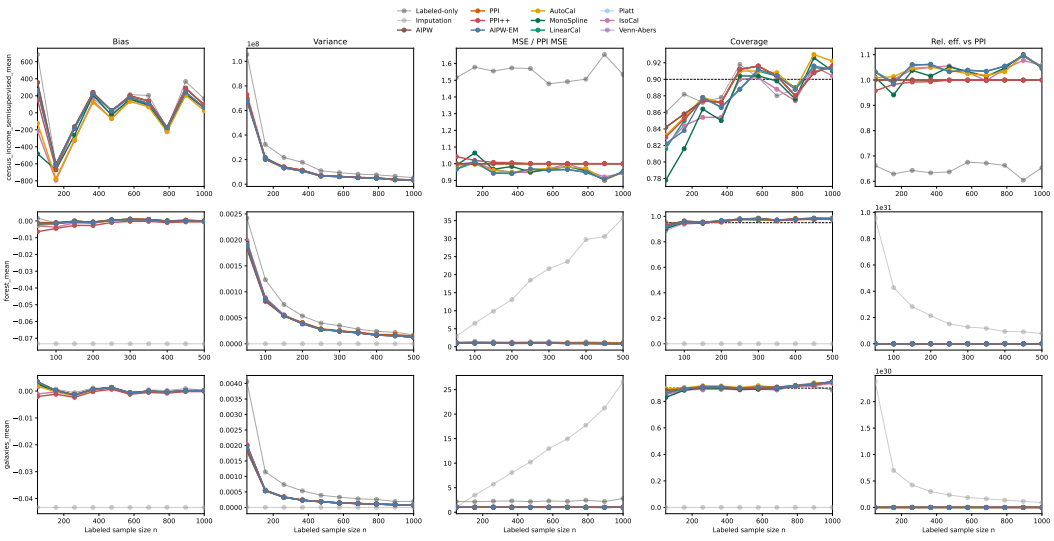

Figure 6: Full diagnostic grid for the reproduced PPI benchmarks, showing bias, empirical variance, normalized MSE relative to PPI, coverage, and relative efficiency versus PPI across labeled-sample-size regimes. In the normalized-MSE and relative-efficiency panels, the dashed horizontal line marks parity with PPI; in the coverage panels it marks the nominal target coverage $1 - \alpha$. `AutoCal`, `MonoSpline`, and `IsoCal` appear alongside PPI, AIPW, PPI++, AIPW-EM, and the fixed calibration rules, while Platt scaling and Venn–Abers appear only on the binary-outcome datasets where they are well defined.

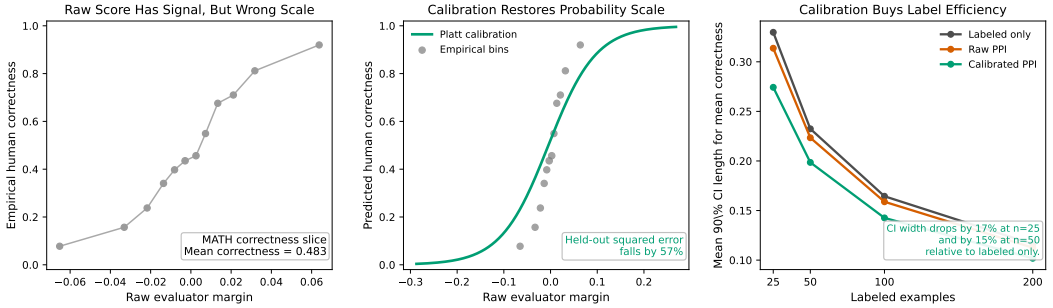

Figure 7: ArmoRM scores and observed MATH correctness. The calibration map is fitted on 100 labeled observations; prediction error is evaluated on the remaining observations.

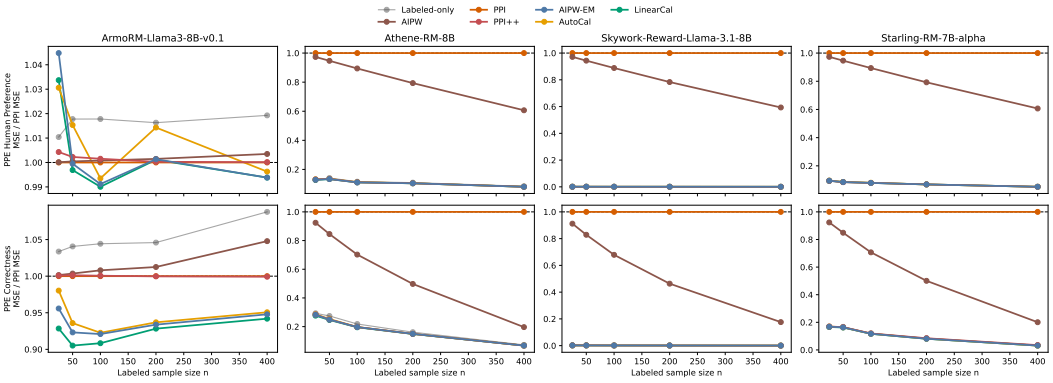

Figure 8: Appendix evaluator-specific PPE summary. Each panel repeats the PPE Human and PPE Correctness comparisons for one public evaluator. The largest MSE increases relative to labeled-only estimation occur for Skywork and Athene.

Track	n	Estimator	MSE / PPI	Label savings	Coverage
PPE Correctness	100	Labeled-only	0.346	0.000	0.927
PPE Correctness	100	PPI	1.000	0.010	0.904
PPE Correctness	100	AIPW	0.774	0.009	0.906
PPE Correctness	100	PPI++	0.330	0.070	0.929
PPE Correctness	100	AIPW-EM	0.309	0.095	0.930
PPE Correctness	100	LinearCal	0.305	0.101	0.931
PPE Correctness	100	AutoCal	0.310	0.094	0.929
PPE Correctness	100	MonoSpline	0.309	0.092	0.926
PPE Correctness	100	IsoCal	0.309	0.089	0.925
PPE Correctness	100	Platt	0.304	0.102	0.930
PPE Correctness	400	Labeled-only	0.299	0.000	0.988
PPE Correctness	400	PPI	1.000	0.020	0.924
PPE Correctness	400	AIPW	0.405	0.009	0.941
PPE Correctness	400	PPI++	0.276	0.083	0.989
PPE Correctness	400	AIPW-EM	0.262	0.103	0.989
PPE Correctness	400	LinearCal	0.260	0.105	0.989
PPE Correctness	400	AutoCal	0.263	0.102	0.988
PPE Correctness	400	MonoSpline	0.263	0.094	0.989
PPE Correctness	400	IsoCal	0.265	0.089	0.989
PPE Correctness	400	Platt	0.260	0.105	0.989
PPE Human Preference	100	Labeled-only	0.303	0.000	0.911
PPE Human Preference	100	PPI	1.000	0.004	0.897
PPE Human Preference	100	AIPW	0.919	0.004	0.896
PPE Human Preference	100	PPI++	0.298	0.027	0.912
PPE Human Preference	100	AIPW-EM	0.295	0.031	0.909
PPE Human Preference	100	LinearCal	0.295	0.031	0.909
PPE Human Preference	100	AutoCal	0.297	0.028	0.908
PPE Human Preference	100	MonoSpline	0.297	0.025	0.903
PPE Human Preference	100	IsoCal	0.297	0.024	0.905
PPE Human Preference	400	Labeled-only	0.289	0.000	0.945
PPE Human Preference	400	PPI	1.000	0.005	0.915
PPE Human Preference	400	AIPW	0.703	0.004	0.917
PPE Human Preference	400	PPI++	0.283	0.025	0.946
PPE Human Preference	400	AIPW-EM	0.282	0.028	0.945
PPE Human Preference	400	LinearCal	0.282	0.028	0.944
PPE Human Preference	400	AutoCal	0.282	0.027	0.944
PPE Human Preference	400	MonoSpline	0.282	0.027	0.942
PPE Human Preference	400	IsoCal	0.282	0.026	0.940

Table 3: PPE numerical summary at $n \in \{100, 400\}$, including the efficiency-maximized and additional fixed-calibration estimators.

n	Estimator	Spearman	Top-1	Top-1 regret
25	AIPW	0.268	0.216	0.064
25	AIPW-EM	0.599	0.344	0.028
25	AutoCal	0.600	0.343	0.028
25	Labeled-only	0.606	0.315	0.030
25	LinearCal	0.599	0.344	0.028
25	PPI	0.265	0.216	0.064
25	PPI++	0.594	0.336	0.029
100	AIPW	0.468	0.285	0.044
100	AIPW-EM	0.845	0.477	0.007
100	AutoCal	0.845	0.487	0.007
100	Labeled-only	0.846	0.485	0.007
100	LinearCal	0.846	0.482	0.007
100	PPI	0.456	0.279	0.046
100	PPI++	0.844	0.479	0.007
400	AIPW	0.631	0.439	0.027
400	AIPW-EM	0.941	0.644	0.002
400	AutoCal	0.941	0.640	0.002
400	Labeled-only	0.940	0.625	0.002
400	LinearCal	0.941	0.646	0.002
400	PPI	0.579	0.419	0.031
400	PPI++	0.940	0.636	0.002

Table 4: Appendix PPE-ranking summary at selected labeled sample sizes from the LLM benchmark. Better rankings have higher Spearman correlation and top-1 identification rates and lower top-1 regret.

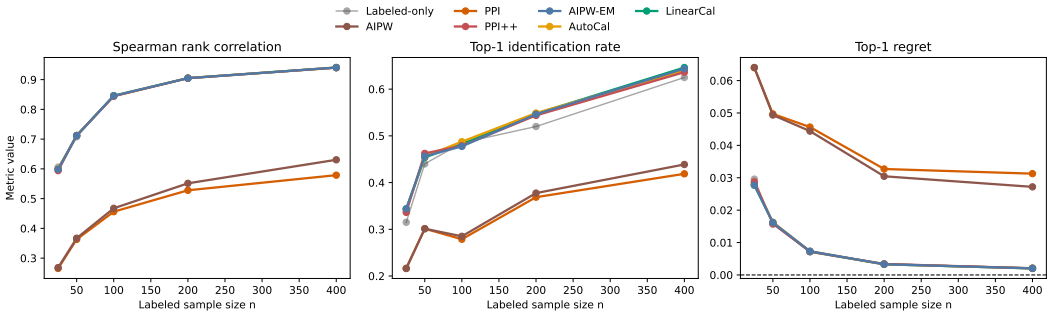

Figure 9: Appendix PPE-ranking supplement from the LLM benchmark, macro-averaged across the public evaluator models. The three panels report Spearman rank correlation, top-1 identification rate, and top-1 regret for recovering the ranking of the eight selected PPE target models from a small labeled sample plus evaluator scores.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main claims, and the paper distinguishes exact representation results, asymptotic theory, and empirical findings; see sections 2 to 5.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Claims in the abstract and introduction should match the paper's contributions and scope.
   - Important assumptions and limitations should be clearly stated.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The theory assumes independent labeled and unlabeled samples, a fixed prediction score, and the stated boundedness and tail conditions. The post-processing result also assumes bounded variation of the conditional mean given the initial score; see sections 4 and D.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Authors are encouraged to discuss strong assumptions, robustness to violations of those assumptions, and the practical scope of the claims.
   - The answer n/a means the paper has no limitations; the answer no means limitations exist but are not discussed.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [Yes].

   Justification: The main assumptions are stated in sections 2 and 4, with complete proofs collected in sections H to J, L and M.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - All assumptions should be stated or referenced in theorem and lemma statements.
   - Complete proofs may appear in the main paper or appendix; informal proof sketches in the main text should be backed by formal proofs.
   - Theorems, lemmas, and external results used in proofs should be properly referenced.

4. **Experimental result reproducibility**

   Question: If the contribution is a dataset, model, algorithm, or empirical evaluation, what steps were taken to make the results reproducible or verifiable?

   Answer: [Yes].

   Justification: The main experimental takeaways appear in section 5, with benchmark construction, additional diagnostics, evaluator-specific LLM results, ranking results, and compute details in sections N and N.3 to N.5.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - Reproducibility can be supported by detailed algorithm descriptions, instructions, code, data, checkpoints, hosted access, or other appropriate avenues.
   - If reproducibility is limited, the paper should explain the available path for verifying or reproducing the main claims.

5. **Open access to data and code**

   Question: If the paper includes experiments, does it include the code, data, and instructions needed to reproduce the main experimental results, either in the supplemental material or as a URL?

Answer: [Yes].

Justification: The experiments use public benchmark sources cited in the paper, and the submission includes implementation sketches and can be accompanied by a supplementary artifact. See sections N.5 and O.

Guidelines:

- The answer should be yes, no, or n/a.
- Instructions should include the exact commands and environment needed to reproduce the main experimental results, including the new method and baselines.
- If only a subset of experiments is reproducible, the paper should say which subset.
- At submission time, released links or archives should preserve anonymity.

6. **Experimental setting/details**

   Question: If the paper includes experiments, does it specify all training and test details, including data splits, hyperparameters, and how they were chosen?

   Answer: [Yes].

   Justification: The simulation design, benchmark protocols, estimator sets, repetition counts, evaluator sources, and summary metrics are described in sections 5, N and N.5.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - Important experimental details should appear in the main paper; full details may also appear in the appendix, code, or supplemental material.
   - Hyperparameter choices and selection procedures should be stated.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The experiments report Monte Carlo variability, MSE, interval coverage, and efficiency comparisons over repeated random splits, and the inference procedures are tied to the Wald theory in section 4; see also corollary 1.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - Error bars, confidence intervals, or significance tests should be reported for experiments supporting the main claims when appropriate.
   - The paper should state what source of variability is shown and how error bars or intervals are computed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Hardware, software versions, repetition counts, use of precomputed evaluator scores, and elapsed time for the small-sample experiment are given in section N.5.

   Guidelines:

   - The answer should be yes, no, or n/a.
   - The answer n/a means the paper does not include experiments.
   - The paper should describe the compute type, memory or hardware when relevant, per-run compute, and estimated total compute.
   - The paper should disclose whether substantially more compute was used for the project than for the reported experiments.

9. **Code of ethics**

   Question: Have the authors read the NeurIPS Code of Ethics and ensured that the research conforms to it?

   Answer: [Yes].

Justification: The paper is a methodological contribution in semisupervised statistical inference, and we are not aware of any aspect of the work that conflicts with the NeurIPS Code of Ethics.

Guidelines:

- The answer should be yes or no, with explanation if needed.
- Any special circumstances requiring deviation from the NeurIPS Code of Ethics should be explained while preserving anonymity.

10. **Broader impacts**

Question: If appropriate for the scope and focus of the paper, does it discuss potential negative societal impacts of the work performed?

Answer: [N/A].

Justification: The paper develops estimation and inference methodology rather than a deployed decision system.

Guidelines:

- The answer should be yes, no, or n/a.
- Authors should discuss plausible negative impacts when there is a direct path to misuse, fairness, privacy, security, or other societal harms.
- Foundational or generic methods may answer n/a when no direct negative application is apparent, but limitations and deployment risks should still be handled honestly.
- Mitigations may be discussed when negative impacts are identified.

11. **Safeguards**

Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

Answer: [N/A].

Justification: The paper does not introduce a system that generates content or interacts with users, and it does not release new high-risk artifacts requiring special safeguards.

Guidelines:

- The answer should be yes, no, or n/a.
- High-risk or dual-use model releases should include appropriate access, usage, or release safeguards.
- Dataset releases that may contain unsafe content should describe how unsafe material was avoided or controlled.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No].

Justification: The sources are publicly available and cited, but the license and usage terms for some precomputed PPE results could not be verified.

Guidelines:

- The answer should be yes, no, or n/a.
- Existing code, data, models, and benchmarks should be cited and their licenses or terms of use should be respected.
- Versions and URLs should be provided when possible.
- If releasing derived assets, the original and derived licenses should be stated where applicable.

13. **New assets**

Question: If new assets are released, are they documented and are relevant details provided alongside the assets?

Answer: [No].

Justification: The reproduction code is documented, but the release does not include a code license.

Guidelines:

- The answer should be yes, no, or n/a.

- The answer n/a means the paper does not release new assets.
- New datasets, code, or models should be documented with training, license, limitations, and intended-use details as relevant.
- Released assets should be anonymized for initial submission when applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper analyzes existing public benchmark labels and does not conduct new human-subject experiments or crowdsourcing.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - The full participant instructions, screenshots, and compensation details should be included when applicable, with more detail in the main paper when human subjects are central to the contribution.
    - Work involving data collection, curation, or other labor should follow the NeurIPS Code of Ethics, including compensation requirements.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: If human-subject research was conducted, does the paper describe participant risks and whether IRB or equivalent approval was obtained?

    Answer: [N/A].

    Justification: The paper uses existing public benchmark data and does not involve new crowdsourcing or human-subject research.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - IRB or equivalent review should be described when required by the relevant institution or jurisdiction.
    - Initial submissions should not include information that breaks anonymity, such as the reviewing institution, when applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes].

    Justification: The LLM evaluation experiments use public, precomputed model scores, as described in Section 5.

    Guidelines:

    - The answer should be yes, no, or n/a.
    - The paper should describe LLM use that is important, original, or non-standard for the core method, experiments, evaluation, or data generation.
    - The answer n/a means LLMs are not important, original, or non-standard components of the core method.
    - LLM use only for writing, editing, or formatting does not need to be declared when it does not affect methodology, rigor, or originality.