Unified Approach for Weakly Supervised Multicalibration

Futoshi Futami[1,2,3] **Takashi Ishida**[2,3]
[1]The University of Osaka [2]The University of Tokyo [3]RIKEN AIP
futami.futoshi.es@osaka-u.ac.jp ishi@k.u-tokyo.ac.jp

Abstract

Multicalibration requires predicted scores to agree with label probabilities across rich families of subgroups and score-dependent tests, but existing methods require clean input–label pairs for evaluation and post-processing. This assumption fails in weakly supervised learning (WSL) regimes—including positive-unlabeled, unlabeled-unlabeled, and positive-confidence learning—where clean labels are costly or unavailable even though reliable uncertainty estimates may be crucial. We address this gap by developing estimators of multicalibration error and post-hoc correction methods for WSL settings in which clean input–label pairs are unavailable. We propose a unified framework for estimating and correcting multicalibration under weak supervision by combining contamination-matrix risk rewrites with witness-based calibration constraints, yielding corrected multicalibration moments with finite-sample guarantees. We further propose weak-label multicalibration boost (WLMC), a generic post-hoc recalibration algorithm under weak supervision. Finally, we conduct experiments across multiple weak-supervision settings to evaluate multicalibration behavior and offer empirical insight into uncertainty estimation under weak supervision.

1 Introduction

Reliable uncertainty estimates are now an indispensable requirement in high-stakes machine learning applications. In domains such as medicine, science, and decision support, a predictor should not only be accurate on average but also output scores that can be trusted as probabilities. Calibration formalizes this requirement by asking predicted probabilities to agree with outcome frequencies [7, 13]. Multicalibration strengthens the requirement by enforcing calibration simultaneously across a rich family of subpopulations and score tests [23, 17, 16]. This viewpoint has made calibration useful far beyond aggregate reliability: it is used to reason about subgroup fairness, distributional reliability, and downstream decisions based on predicted probabilities [11].

Despite this progress, existing multicalibration methods still face a basic limitation. To the best of our knowledge, they assume access to complete input-label pairs both when computing calibration metrics and when running post-processing. In some applications, however, clean labels are expensive to obtain or unavailable at scale, while only weaker supervision signals are observed. Such situations have motivated weakly supervised learning (WSL), especially for classification, where predictors can be learned even without complete label information. For example, in positive-unlabeled (PU) learning [10, 36, 28], only positive and unlabeled data may be available; other representative examples include unlabeled-unlabeled (UU) learning [31, 32] and positive-confidence (Pconf) learning [25].

In these settings, weak supervision creates a basic asymmetry: it can provide enough information to train an accurate classifier, while withholding the clean label-frequency information needed to evaluate or correct calibration directly. This is problematic because WSL is often used where clean labels are costly, delayed, or unavailable—including medicine and decision support—yet learned scores

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

set thresholds, prioritize cases, and guide downstream actions. Thus, the settings in which reliable uncertainty estimates are most needed are often the same settings in which multicalibration and post-processing must be performed using only weakly supervised data. Since calibration is defined through label frequencies, and multicalibration through subgroup- and score-dependent residual moments, standard metrics and post-processing methods cannot be applied without modification.

Only limited work addresses this obstacle. Most closely, Kiryo et al. [29] extend binned expected calibration error (ECE) to PU learning. Their estimator is derived directly for the PU binned-ECE setting by applying Bayes' rule within each score bin. This result is important, but it addresses only PU binned ECE: it does not cover multicalibration over rich witness classes, does not give a post-processing method, and does not address other weak-observation models. More broadly, weakly supervised learning contains many observation models beyond PU learning (see Appendix B), and deriving a separate metric and correction algorithm for each model would fragment the theory, complicate comparison and model selection, and obscure the relationships among settings.

This motivates a unified viewpoint. A natural starting point is the unified risk-analysis view of weakly supervised learning [5]. There, weak supervision is described by a contamination matrix, or more generally an operator, that maps inaccessible clean distributions to observable weak distributions; target clean risks are then recovered by a decontamination rewrite. A key advantage is that, once the observable marginal information and the contamination mechanism are specified, estimators whose expectations equal the target losses can be derived in a systematic way.

In this paper, we use the WSL risk-rewrite principle in a different statistical role. Decontamination identities in WSL are most commonly used to construct empirical risks for training classifiers from weak observations. Here, the predictor is fixed, and the target is not a training objective but a calibration residual that measures whether predicted probabilities agree with labels after restricting attention to a subgroup and a score-dependent test. This shift matters for multicalibration: the goal is not to optimize a single corrected risk, but to audit a family of residuals and control the worst violation over the witness class. For each fixed witness, the residual can be made observable by the same decontamination machinery as an ordinary loss; the remaining statistical question is whether the corrected empirical residuals can be controlled uniformly and used to drive post-hoc correction.

Our contributions are as follows. First, we introduce a witness-level decontamination reduction for multicalibration (Theorem 1), which treats calibration constraints as signed residual auditing functionals rather than as training losses. This yields corrected PU, UU, and Pconf estimators (Corollaries 1, 2, and 3) and finite-sample uniform guarantees for weak multicalibration estimates. Second, we use these moments in weak-label multicalibration boost (WLMC), a weak-label post-processing algorithm, and prove its finite-sample guarantee in Theorem 2. Finally, to the best of our knowledge, we give the first real-data study of weak-label multicalibration, using weak labels for both estimation and post-hoc correction. The experiments yield three takeaways: weak multicalibration estimates can track oracle estimates; weakly trained predictors are often less calibrated; and weak labels can improve post-hoc calibration even for clean-label predictors, suggesting a lower-cost route for improving multicalibration when clean calibration labels are expensive.

The paper is organized as follows. Section 2 discusses related work, Section 3 introduces multicalibration and weak supervision, Section 4 introduces our corrected metrics, Section 5 proposes the post-processing method, and Section 6 reports the empirical study.

2 Related work

Multicalibration was introduced by Hebert-Johnson et al. [23] as a subgroup fairness notion, and has since been studied through efficient algorithms, boosting, and statistical guarantees, and empirical evaluations [17, 16, 18, 21]. The broader calibration literature studies classical forecasting calibration, neural-network calibration, smooth or IPM-style metrics, trainable calibration objectives, scalable testing, and adaptive evaluation [7, 13, 19, 2, 40, 33, 24, 14]. Existing multicalibration and decision-calibration style guarantees, however, are developed under clean labels. Our work is complementary: we study how to estimate and correct calibration witnesses when only weak labels are available.

Our technical starting point is the decontamination view of WSL, which represents weak-supervision models through contamination matrices and rewrites clean risks from weak data [5]. These rewrites are most commonly used to build empirical objectives for classifier learning. We instead apply the same principle to witness-based calibration residuals and post-hoc correction. The closest prior

calibration-estimation work is PU-ECE [29], which estimates binned ECE by applying Bayes' rule within each score bin; in our framework, this bin correction corresponds to using a constant subgroup function equal to one and a binning function for the score weight. Our framework gives a different abstraction: each calibration constraint is a signed residual moment indexed by a witness consisting of a subgroup function and a score-weighting function, and the weak-observation model enters only through the decontamination operator. Our results cover subgroup-dependent witnesses, UU and Pconf supervision, uniform convergence over subgroup and score-test classes, and post-hoc correction. Together, these results provide a first step toward multicalibration guarantees in WSL.

A related line of work studies uncertainty quantification when clean labels are unavailable or corrupted. Conformal prediction has been extended to weak supervision and analyzed under label noise, while PU-ECE studies calibration-error estimation in positive-unlabeled learning [4, 12, 29]. These works show that uncertainty guarantees must account for the label observation process. Our work follows this motivation, but focuses on multicalibration and post-hoc correction in WSL.

3 Preliminaries

3.1 Multicalibration for binary classification

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces. Throughout, we focus on binary classification, $\mathcal{Y} = \{0, 1\}$. Let $P(Y, X)$ be the joint distribution over $\mathcal{X} \times \mathcal{Y}$, and write the class priors as $\pi_+ = P(Y = 1)$ and $\pi_- = P(Y = 0)$. We denote the class-conditional distributions by $P_+ = P(X \mid Y = 1)$ and $P_- = P(X \mid Y = 0)$, and the regression function by $r(x) = P(Y = 1 \mid X = x)$. Following notation used in weakly supervised learning [5], we also use the compact joint-measure vector

$$P := (P(X, Y = 1), P(X, Y = 0))^\top = (\pi_+ P_+, \pi_- P_-)^\top.$$

A predictor is a measurable function $f : \mathcal{X} \to [0, 1]$, intended to estimate the conditional class probability $f^\star(x) := \mathbb{E}[Y \mid X = x]$.

Let $\mathcal{C} \subseteq \{c : \mathcal{X} \to [0, 1]\}$ be a class of subgroup functions, and let $\mathcal{W} \subseteq \{w : [0, 1] \to \mathbb{R} : \|w\|_\infty \leq 1\}$ be a class of score-dependent test functions. We use the following witness-based definition of multicalibration from Gopalan et al. [17].

Definition 1 (Multicalibration)**.** *For a predictor f, define its $(\mathcal{C}, \mathcal{W})$-multicalibration error by*

$$\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left| \mathbb{E}\big[c(X) w(f(X))(Y - f(X))\big] \right|. \tag{1}$$

We say that f is $(\mathcal{C}, \mathcal{W}, \varepsilon)$-multicalibrated if $\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f) \leq \varepsilon$.

This definition subsumes familiar calibration criteria by specifying the witness classes $\mathcal{C}$ and $\mathcal{W}$. With $\mathcal{C} = \{1\}$, taking $\mathcal{W}$ to be all measurable maps $[0, 1] \to [-1, 1]$ recovers ECE, while $\mathcal{W} = \mathrm{Lip}_1([0, 1], [-1, 1])$ gives smooth calibration error [2]. Binned ECE is obtained by fixing $\delta \in (0, 1]$, setting $J = \lceil 1/\delta \rceil$, $I_j = [(j-1)\delta, j\delta)$ for $j < J$ and $I_J = [(J-1)\delta, 1]$, and using $\mathcal{W}_\delta = \{p \mapsto \mathbf{1}\{p \in I_j\} : j = 1, \ldots, J\}$; with a nontrivial $\mathcal{C}$, this is the usual binned form of multicalibration [20, 23]. The class $\mathcal{C}$ specifies the groups or tests on which calibration must hold. Indicator classes $\mathcal{C} = \{\mathbf{1}_G : G \in \mathcal{G}\}$ give subgroup fairness notions [23, 26]; richer choices such as decision trees, or linear thresholds require the calibration residual to be indistinguishable from zero against a broad test class, connecting multicalibration and multiaccuracy [27, 11]. Given independent and identically distributed (i.i.d.) samples $S = \{(X_i, Y_i)\}_{i=1}^n$ from $P(X, Y)$, the empirical analogue is

$$\widehat{\mathrm{MC}}_{\mathcal{C},\mathcal{W}}(f; S) := \sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left| \frac{1}{n} \sum_{i=1}^{n} c(X_i) w(f(X_i))(Y_i - f(X_i)) \right|.$$

3.2 Weakly supervised learning and unified risk rewrites

WSL replaces clean input–label pairs (X, Y) with weaker observations. In PU learning [10, 36, 28, 29], the sample consists of independent positive and unlabeled examples,

$$S_+ = \{X_i^+\}_{i=1}^{n_+} \overset{\text{i.i.d.}}{\sim} P_+, \qquad S_u = \{X_j^u\}_{j=1}^{n_u} \overset{\text{i.i.d.}}{\sim} P_X, \qquad P_X = \pi_+ P_+ + \pi_- P_-. \tag{2}$$

For any binary loss $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}$, the clean risk admits the PU rewrite

$$\begin{aligned} R_\ell(f) = \mathbb{E}_{P(X,Y)}[\ell(f(X),Y)] &= \pi_+ \mathbb{E}_{P_+}[\ell(f(X),1)] + \pi_- \mathbb{E}_{P_-}[\ell(f(X),0)] \\ &= \pi_+ \mathbb{E}_{P_+}[\ell(f(X),1) - \ell(f(X),0)] + \mathbb{E}_{P_X}[\ell(f(X),0)]. \end{aligned} \tag{3}$$

This yields the corrected empirical estimator

$$\widehat{R}_\ell^{\mathrm{PU}}(f) = \frac{\pi_+}{n_+} \sum_{i=1}^{n_+} \{\ell(f(X_i^+),1) - \ell(f(X_i^+),0)\} + \frac{1}{n_u} \sum_{j=1}^{n_u} \ell(f(X_j^u),0). \tag{4}$$

This example highlights the central WSL principle: rather than recovering the hidden label for each instance, we rewrite the target expectation in terms of quantities observable under the weak sampling scheme. Throughout our PU analysis, we treat π_+ as known or estimated in a preprocessing step, as is standard in PU learning [10, 36, 28, 29].

The unified framework of Chiang & Sugiyama [5] expresses such rewrites systematically in matrix form. Set $B := (P_+, P_-)^\top$, $\Pi := \mathrm{diag}(\pi_+, \pi_-)$, and $P = \Pi B = (\pi_+ P_+, \pi_- P_-)^\top$.

For a binary loss, define $L_\ell(x; f) := (\pi_+ \ell(f(x),1), \pi_- \ell(f(x),0))^\top$, so that

$$R_\ell(f) = \int_{\mathcal{X}} L_\ell(x; f)^\top \, dB(x). \tag{5}$$

A weak-observation model specifies an observable source distribution $\bar{P}$ and a decontamination rule

$$\bar{P} = M_{\mathrm{corr}} B, \qquad B = M_{\mathrm{corr}}^\dagger \bar{P}, \tag{6}$$

where $M_{\mathrm{corr}}^\dagger$ may be an inverse matrix or an x-dependent operator. This relation gives the risk rewrite

$$R_\ell(f) = \int_{\mathcal{X}} L_\ell(x; f)^\top M_{\mathrm{corr}}^\dagger \, d\bar{P}(x). \tag{7}$$

For PU, the identities used below are

$$\bar{P}_{\mathrm{PU}} := (P_+, P_X)^\top = M_{\mathrm{PU}} B, \qquad M_{\mathrm{PU}} := \begin{pmatrix} 1 & 0 \\ \pi_+ & \pi_- \end{pmatrix}, \tag{8}$$

$$M_{\mathrm{PU}}^\dagger = M_{\mathrm{PU}}^{-1} = \begin{pmatrix} 1 & 0 \\ -\pi_+/\pi_- & 1/\pi_- \end{pmatrix}, \qquad B = M_{\mathrm{PU}}^\dagger \bar{P}_{\mathrm{PU}}. \tag{9}$$

Plugging (9) into (7), we recover Eqs. (3) and (4). Other weak-observation models use the same template with a different M_{corr}; Appendix B lists the variants used later.

4 Weakly supervised multicalibration

Multicalibration under weak supervision faces two difficulties: the residual $Y - f(X)$ is label-dependent although labels are observed only indirectly, and the error is a supremum over subgroup- and score-dependent witnesses rather than a single risk. Our key perspective is to regard each multicalibration witness moment as a signed target risk, not as a loss to be minimized for training but as an auditing functional for a fixed probabilistic predictor. This fixed-witness view lets us apply the same decontamination principle as in weakly supervised risk estimation; the remaining task is to control the corrected moments uniformly over the witness class and use them for post-processing.

We develop this idea in two steps. Section 4.1 gives the generic decontamination rewrite, and Section 4.2 instantiates it for PU, UU, and Pconf learning. We state and prove the finite-sample guarantee in detail for PU in the main text; analogous guarantees for UU and Pconf follow under their corresponding envelope and sampling assumptions.

4.1 Unified rewrite of multicalibration in the WSL setting

Section 3.2 showed how clean risks can be rewritten under weak supervision. We now lift the same algebra from ordinary losses to multicalibration witnesses.

For any fixed $(c, w) \in \mathcal{C} \times \mathcal{W}$, define

$$\phi_f^{c,w}(x,y) := c(x) w(f(x))(y - f(x)), \qquad R_{c,w}(f) := \mathbb{E}\big[\phi_f^{c,w}(X,Y)\big].$$

Then (1) becomes $\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f) = \sup_{c,w} |R_{c,w}(f)|$. Similarly to Eq. (5), conditioning on Y gives

$$L_{c,w}(x; f) := \begin{pmatrix} \pi_+ c(x) w(f(x))(1 - f(x)) \\ -\pi_- c(x) w(f(x)) f(x) \end{pmatrix}, \qquad R_{c,w}(f) = \int L_{c,w}(x; f)^\top \, dB(x).$$

Similarly to Eq. (7), we obtain the following rewrite of the multicalibration moment.

Theorem 1 (Generic WSL rewrite for multicalibration)**.** *Assume that the weak-supervision model admits a decontamination matrix or operator $M^\dagger_{\mathrm{corr}}$ such that $B = M^\dagger_{\mathrm{corr}} \bar{P}$ as a signed-measure identity, and that the corrected integrand is integrable. For any fixed witness (c, w), define*

$$R^{\mathrm{WSL}}_{c,w}(f) := \int_{\mathcal{X}} L_{c,w}(x; f)^\top M^\dagger_{\mathrm{corr}} \, d\bar{P}(x). \tag{10}$$

Then $R_{c,w}(f) = R^{\mathrm{WSL}}_{c,w}(f)$ holds. Consequently, multicalibration admits the representation

$$\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f) = \sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left| R^{\mathrm{WSL}}_{c,w}(f) \right|, \tag{11}$$

The proof is given in Appendix A. For a fixed witness, the equality follows from the same decontamination principle used in weakly supervised risk estimation. The difference is its statistical role: instead of constructing an empirical objective for training a classifier, we use the rewrite to make calibration residual moments observable for a fixed probabilistic predictor. Multicalibration then requires controlling not one corrected risk, but a witness-indexed signed residual process over $C \times W$. The finite-sample results below show that these corrected moments can be controlled uniformly and therefore used as weak-label auditors for post-hoc correction.

Let $\widehat{R}^{\mathrm{WSL}}_{c,w}(f; S)$ denote the finite-sample estimator of $R^{\mathrm{WSL}}_{c,w}(f)$ on a weak dataset S. The empirical estimator is obtained by replacing the relevant expectations with the corresponding source-wise empirical averages for the data-generation mechanism. Section 4.2 writes these estimators explicitly.

4.2 Three concrete weak-supervision examples

We instantiate Theorem 1 in three binary settings: PU, UU, and Pconf. These examples show how the multicalibration moment is corrected under weak supervision. Other weak-supervision specifications, including SU, DU, SD, pairwise comparison, soft-label, and Sconf variants, are recorded in Appendix B. To keep the main text focused, we state the finite-sample guarantee for PU; analogous arguments for UU and Pconf are provided in Appendix C.

PU learning. We reuse the PU sample model and matrices from Section 3.2, namely Eq. (2) and Eq. (8)–(9). Substituting $M^\dagger_{\mathrm{PU}}$ into the generic rewrite in Eq. (10) gives the following expression.

Corollary 1. *Under the PU observation model, for every fixed $(c, w) \in \mathcal{C} \times \mathcal{W}$,*

$$R^{\mathrm{PU}}_{c,w}(f) = \pi_+ \mathbb{E}_{P_+}[c(X) w(f(X))] - \mathbb{E}_{P_X}[c(X) w(f(X)) f(X)].$$

Given $\{X^+_i\}_{i=1}^{n_+} \overset{\text{i.i.d.}}{\sim} P_+$ and $\{X^u_j\}_{j=1}^{n_u} \overset{\text{i.i.d.}}{\sim} P_X$, the corresponding finite-sample estimator is

$$\widehat{R}^{\mathrm{PU}}_{c,w}(f) = \frac{\pi_+}{n_+} \sum_{i=1}^{n_+} c(X^+_i) w(f(X^+_i)) - \frac{1}{n_u} \sum_{j=1}^{n_u} c(X^u_j) w(f(X^u_j)) f(X^u_j). \tag{12}$$

The multicalibration error estimator is $\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c,w} |\widehat{R}^{\mathrm{PU}}_{c,w}(f)|$.

The following proposition controls the finite-sample deviation of this estimator.

Proposition 1 (Uniform convergence for the PU weak estimate)**.** *Assume that f is fixed independently of the PU samples. Define $(P_1, n_1, \pi_1) := (P_+, n_+, \pi_+)$ and $(P_2, n_2, \pi_2) := (P_X, n_u, 1)$. Then for any $\delta \in (0, 1]$ and $\eta \in (0, 1]$, with probability at least $1 - \delta$,*

$$\left| \widehat{\mathrm{MC}}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}_{\mathcal{C},\mathcal{W}}(f) \right| \le 2 \sum_{s=1}^{2} \pi_s \left(\mathfrak{R}_{n_s}(\mathcal{C}; P_s) + \eta + \sqrt{\frac{2 \log N_\infty(\eta, \mathcal{W}) + \log(4/\delta)}{n_s}} \right), \tag{13}$$

where $\mathfrak{R}_n(\mathcal{C}; P) := \mathbb{E}_{S \sim P^n, \sigma}\left[\sup_{c \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i c(X_i) \right]$ is the expected Rademacher complexity of $\mathcal{C}$ under P, $\sigma_i \in \{-1, +1\}$ are independent Rademacher random variables, and $N_\infty(\eta, \mathcal{W}) = N(\eta, \mathcal{W}, \|\cdot\|_\infty)$ is the covering number of $\mathcal{W}$ under the sup norm.

This PU bound shows that both positive and unlabeled samples are needed to control finite-sample estimation error. It also yields sample-complexity rates once $\mathcal{C}$ and $\mathcal{W}$ are specified: for example, when $n = n_+ = n_u$ and $\mathcal{C}$ and $\mathcal{W}$ are finite, as in standard binning-based MC, $\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}_{\mathcal{C},\mathcal{W}}(f)\right| = \mathcal{O}_p(\sqrt{\log(2|\mathcal{C}||\mathcal{W}|)/n})$. This logarithmic dependence is important: even though the estimator takes a worst case over subgroup and score witnesses, the price of uniformly estimating corrected moments grows logarithmically with the sizes of group and scoring classes. Hence, once enough samples are available from each observed source, the weak-label estimator remains efficient. The proof and bounds for other choices of $\mathcal{C}$ and $\mathcal{W}$ are given in Appendix C.2.

UU learning. We now move to the more surprising case where *no* labeled sample is available. In UU learning [31, 32], one observes two unlabeled sources with different mixture proportions

$$P_{U_1} = (1-\gamma_1)P_+ + \gamma_1 P_-, \qquad P_{U_2} = \gamma_2 P_+ + (1-\gamma_2)P_-, \qquad \Delta := 1-\gamma_1-\gamma_2 \neq 0.$$

We assume throughout this subsection that the mixture proportions $\gamma_1, \gamma_2 \in [0,1]$ are fixed and known a priori. Thus PU and UU are both mutually contaminated-distribution (MCD) settings: the observed sources are fixed mixtures of the two class-conditionals, with no x-dependent correction. Equivalently, $M_{\mathrm{UU}} = \left(\begin{smallmatrix} 1-\gamma_1 & \gamma_1 \\ \gamma_2 & 1-\gamma_2 \end{smallmatrix}\right)$, $\bar{P} = \left(\begin{smallmatrix} P_{U_1} \\ P_{U_2} \end{smallmatrix}\right) = M_{\mathrm{UU}} B$, with inverse $M_{\mathrm{UU}}^{-1} = \frac{1}{\Delta}\left(\begin{smallmatrix} 1-\gamma_2 & -\gamma_1 \\ -\gamma_2 & 1-\gamma_1 \end{smallmatrix}\right)$. This uses the same decontamination idea as PU, except that both sources are unlabeled mixtures. Indeed, PU is recovered by setting $(\gamma_1, \gamma_2) = (0, \pi_+)$ with $(P_{U_1}, P_{U_2}) = (P_+, P_X)$. The key point of the UU case is that, although no sample is labeled, the two mixtures are sufficient to estimate the multicalibration error. Substituting M_{UU}^{-1} into the generic rewrite gives the following corrected moment and empirical estimator; the finite-sample bound is given in Appendix C.

Corollary 2. *Under the UU observation model, for every fixed $(c, w) \in \mathcal{C} \times \mathcal{W}$,*

$$R^{\mathrm{UU}}_{c,w}(f) = \sum_{k=1}^{2} \mathbb{E}_{P_{U_k}}[\alpha_k(f(X))c(X)w(f(X))],$$

with $\alpha_1(v) := \frac{(1-\gamma_2)\pi_+(1-v)+\gamma_2\pi_-v}{\Delta}$, $\alpha_2(v) := \frac{-\gamma_1\pi_+(1-v)-(1-\gamma_1)\pi_-v}{\Delta}$. Given samples $\{X_i^{U_1}\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{U_1}$ and $\{X_j^{U_2}\}_{j=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{U_2}$, the corresponding estimator is

$$\widehat{R}^{\mathrm{UU}}_{c,w}(f) = \sum_{k=1}^{2} \frac{1}{n_k} \sum_{i=1}^{n_k} \alpha_k(f(X_i^{U_k}))c(X_i^{U_k})w(f(X_i^{U_k})). \tag{14}$$

Pconf learning. Unlike the PU and UU cases, which use fixed mixture proportions, Pconf [25] learning considers a soft-label setting in which each positive example carries an x-dependent confidence. We observe i.i.d. pairs (X_i, r_i) with $X_i \sim P_+$ and $r_i = r(X_i) = P(Y=1 \mid X_i)$. We treat π_+ as known or pre-estimated and assume the standard support condition $P_- \ll P_+$, so that the negative class can be recovered from positive-confidence data. With $B = (P_+, P_-)^\top$, define the duplicated observable source $\bar{P}_{\mathrm{Pconf}}(dx) := \left(\begin{smallmatrix} P_+(dx) \\ P_+(dx) \end{smallmatrix}\right)$. Bayes' rule gives $P_-(dx) = \frac{\pi_+}{\pi_-}\frac{1-r(x)}{r(x)}P_+(dx)$, and hence the x-dependent decontamination rule $B(dx) = M^{\dagger}_{\mathrm{Pconf}}(x)\bar{P}_{\mathrm{Pconf}}(dx)$ and $M^{\dagger}_{\mathrm{Pconf}}(x) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & \frac{\pi_+}{\pi_-}\frac{1-r(x)}{r(x)} \end{smallmatrix}\right)$. The density-ratio derivation is deferred to Appendix B.3. Substituting this operator into the generic witness rewrite gives the following corrected moment and estimator.

Corollary 3. *For every fixed $(c, w) \in \mathcal{C} \times \mathcal{W}$,*

$$R^{\mathrm{Pconf}}_{c,w}(f) = \pi_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - f(X)r(X)^{-1}\right)\right].$$

With Pconf data $\{(X_i, r_i)\}_{i=1}^{n}$ where $X_i \sim P_+$ and $r_i = r(X_i)$,

$$\widehat{R}^{\mathrm{Pconf}}_{c,w}(f) = \frac{\pi_+}{n}\sum_{i=1}^{n} c(X_i)w(f(X_i))\left(1 - f(X_i)r_i^{-1}\right).$$

The finite-sample estimation error is analyzed in Appendix C.3.

We assume known nuisance parameters, such as π_+ for PU and γ_1, γ_2 for UU. Appendix D discusses how their estimation errors affect the multicalibration estimation error.

Algorithm 1 Finite-sample weak-label multicalibration boost (WLMC)

Input: $\varepsilon > 0$, fixed weak calibration samples S, step size λ, initial predictor f_0, classes $\mathcal{C}, \mathcal{W}$, estimator $\widehat{R}^{\mathrm{WSL}}$
Set $t = 0$
while $\exists(\tau, c, w) \in \{\pm 1\} \times \mathcal{C} \times \mathcal{W}$ such that $\tau \widehat{R}^{\mathrm{WSL}}_{c,w}(f_t; S) > 3\varepsilon/4$ **do**
  Choose $(\tau_t, c_t, w_t) \in \{\pm 1\} \times \mathcal{C} \times \mathcal{W}$ satisfying the while condition
  $\forall x \in \mathcal{X}, \quad f_{t+1}(x) = \Pi_{[0,1]}[f_t(x) + \lambda \tau_t c_t(x) w_t(f_t(x))]$
  where $\Pi_{[0,1]}[z] := \min\{1, \max\{0, z\}\}$ is the projection of z onto $[0, 1]$.
  $t \leftarrow t + 1$
end while
Output: f_t

5 Post-processing methods

This section develops post-processing methods based on the corrected moments in Section 4: WLMC, a weak-supervision multicalibration booster, and weak versions of Platt scaling and temperature scaling that use the same risk rewrites.

5.1 WLMC: weak-label multicalibration boost

Algorithm 1 is our proposed post-processing algorithm. It is a boosting-style recalibration method, in the spirit of Hebert-Johnson et al. [23], Gopalan et al. [17]: at each iteration, it searches for a direction along which multicalibration is violated and moves the predictor in that direction. The search ranges over $(c, w) \in \mathcal{C} \times \mathcal{W}$. The main difference from existing methods is that they assume independent samples at each iteration, i.e., fresh samples, whereas our method reuses the same samples throughout the algorithm. Nevertheless, as shown by the following theorem, the sample complexity remains the same.

Theorem 2 ((Informal) PU guarantee for WLMC)**.** *Let $0 < \varepsilon \leq 1$ and $0 < \delta < 1$. Assume the PU setting of Proposition 1 with known class prior π_+, nonempty finite $\mathcal{C}$ and $\mathcal{W}$, and $n_+ = n_u = n$.*[1] *Assume that $f_0, \mathcal{C}, \mathcal{W}$ are fixed independently of the calibration samples. Let $D_0 = \mathbb{E}[(f_0(X) - f^\star(X))^2]$ and $T = \lceil 4D_0/\varepsilon^2 \rceil$. Draw n i.i.d. samples from each of P_+ and P_X, independently across sources, and run Algorithm 1. For a universal constant $C > 0$, if $n \geq C \log((2|\mathcal{C}||\mathcal{W}|)^{T+2}/\delta)/\varepsilon^2$, then Algorithm 1 with $\lambda = \varepsilon/2$ stops after at most T successful updates and returns a $(\mathcal{C}, \mathcal{W}, \varepsilon)$-multicalibrated hypothesis, both with probability at least $1 - \delta$.*

Therefore, n is the total number of samples from each of the positive and unlabeled sources. Since $D_0 \leq 1$, a sufficient total budget per source is $\mathcal{O}(\log(2|\mathcal{C}||\mathcal{W}|)/\varepsilon^4 + \log(1/\delta)/\varepsilon^2)$. The ε^{-4} dependence matches that of existing methods [23, 17]. UU- and Pconf-based WLMC variants also require the same $O(\varepsilon^{-4})$ sample complexity; see Appendix E with the formal statement. For comparison with prior work, see Section 2.

5.2 Platt scaling and temperature scaling

Recent empirical work on multicalibration [21] reports that classical post-processing methods, such as Platt scaling [35] and temperature scaling [19], can perform competitively with boosting-based methods that come with theoretical guarantees. We therefore implement weak-supervision analogues of these methods. Both keep the base predictor fixed and apply a simple transformation to its output logit. Let $\sigma(t) = (1 + e^{-t})^{-1}$, $\mathrm{logit} = \sigma^{-1}$, and $z_f(x) = \mathrm{logit}(f(x))$. Temperature scaling uses $g_\beta(x) = \sigma(z_f(x)/\beta)$ with parameter $\beta > 0$, while Platt scaling uses $g_{a,b}(x) = \sigma(a z_f(x) + b)$ with parameters $(a, b) \in \mathbb{R}^2$. With clean labels, both are fitted by minimizing held-out negative log-likelihood. Under weak supervision, we keep the same parametric families but replace the clean held-out NLL with corrected weak NLL objectives derived from the risk rewrites in Section 3.2. Appendix E.4 lists the concrete corrected NLL objectives for the three weak-supervision models.

[1] The equal-size assumption is made only for ease of presentation; see Appendix E.2 for the unequal-source-size case.

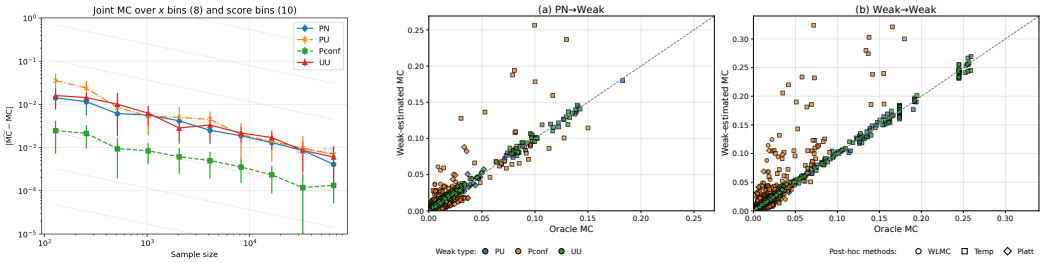


Figure 1: Left: toy-data verification of MC estimators. Middle/right: oracle MC (x-axis; PN labels) versus weak-estimated MC (y-axis) for $(\mathrm{PN}, \mathrm{weak})$ and $(\mathrm{weak}, \mathrm{weak})$ on tabular data. The dashed diagonal denotes exact agreement.

6 Experiments

We empirically study multicalibration under weak supervision on toy and real datasets. Beyond validating the finite-sample estimators in Section 4, the real-data experiments are meant to test whether weak labels can be used as a practical calibration resource: can they estimate multicalibration reliably, and can they support post-hoc correction without requiring an additional clean calibration set? To the best of our knowledge, this is the first numerical study of multicalibration on real data in which the calibration metric and post-processing step are both implemented from weak labels. All main-text results use finite subgroup and prediction-bin collections $\mathcal{C}$ and $\mathcal{W}$; experimental details are deferred to Appendices F–I. For brevity, we refer to multicalibration error as *MC* in this section.

6.1 Toy data: Verification of finite-sample estimation error

We first evaluate the finite-sample estimation error of multicalibration metrics under the WSL settings in Section 4. We generate a toy binary classification problem with a known weak-observation mechanism. We plot the absolute error between each population metric and its estimate as a function of sample size. This experiment verifies the convergence behavior predicted by our analysis in Section 4. All metrics should exhibit an $n^{-1/2}$ rate on a log–log plot. The empirical curves in the left panel of Figure 1 follow this reference rate.

6.2 Real data experiments

Following Hansen et al. [21], we study MC under weak supervision on real datasets. For a baseline predictor f_0, we vary the data for base training and post-hoc correction: $(\mathrm{base}, \mathrm{post}) \in \{(\mathrm{PN}, \mathrm{PN}), (\mathrm{PN}, \mathrm{weak}), (\mathrm{weak}, \mathrm{weak})\}$. PN denotes fully supervised view with complete input–label pairs, while weak denotes a PU, UU, or Pconf. The weak-observation type for base learning and post-processing is matched when both are weak. We call estimates computed with clean labels *oracle estimates (MC)*, and those computed with weak data *weak estimates (MC)*. Weak post-processing uses only the corresponding weak validation criterion for selection; oracle labels are not used for weak fitting or selection. Protocol details are deferred to Appendices G–I. Within each benchmark, PN and weak post-processing use the same correction split; tabular data uses correction fraction 0.4, while CelebA and CivilComments use 0.2. Thus PN and weak post-processing receive comparable correction data, while the weak setting is more stringent because clean labels are withheld.

We summarize the main empirical results in three findings; comprehensive results across supervision settings and post-processing methods are reported in Appendices K and L. **F1.** Weak estimates can proxy oracle estimates when the weak-observation model is well specified. **F2.** Weak-data post-processing can improve PN-trained predictors, suggesting weak labels can serve as possible lower-clean-label-cost calibration signals. **F3.** Weakly trained predictors are less calibrated, but their larger initial violations make weak-data post-processing especially useful. Together, these findings suggest that weak labels may reduce the need for clean calibration labels while still enabling effective multicalibration correction.

Tabular data. We used four tabular datasets—ACS Income, UCI Credit Default, HMDA, and MEPS (see Appendix G for details)—and six base predictors (decision tree, random forest, SVM, naive Bayes, logistic regression, MLP). We first examine whether multicalibration error can be

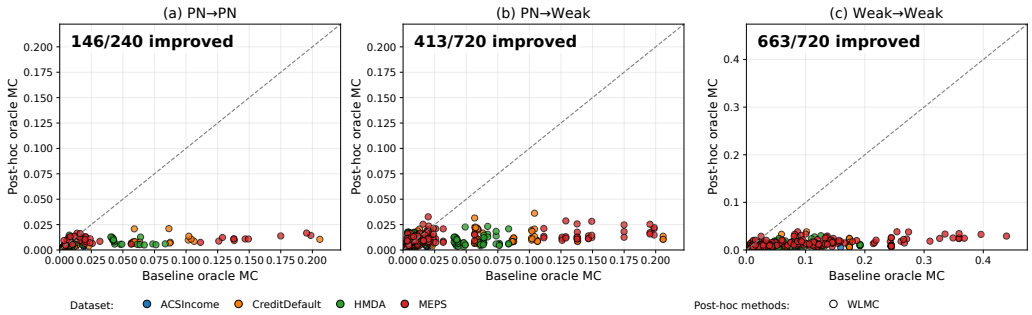


Figure 2: Oracle test MC before and after WLMC.

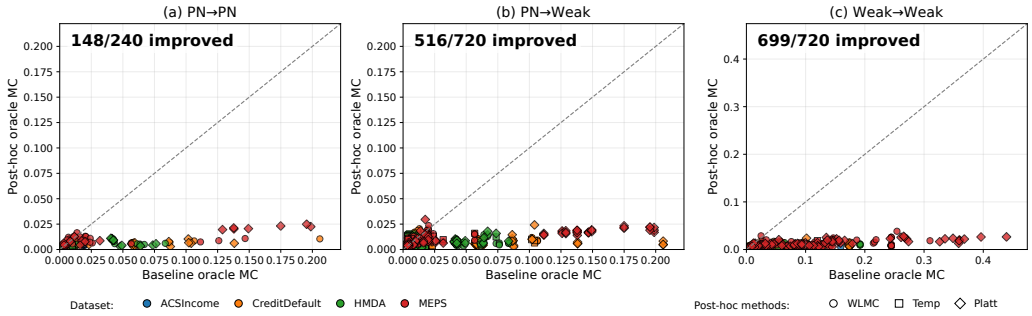


Figure 3: Oracle test MC before and after correction by a method selected from WLMC, temperature scaling, and Platt scaling using the smallest oracle validation MC.

estimated reliably from weak labels. The middle and right panels of Figure 1 compare oracle MC and weak MC across tabular datasets, baselines, weak-label types, and post-hoc correction methods. The diagonal $y = x$ denotes exact agreement: points closer to this line indicate that weak MC more accurately approximates oracle MC. The same plots with correlation coefficients are provided in Appendix K. Overall, weak MC is well aligned with oracle MC. PU and UU points are mostly concentrated near the diagonal, suggesting that weak labels can provide useful MC estimates when the weak-observation model is well matched. Pconf, however, shows several deviations. This is likely because Pconf confidence values are generated from conditional probabilities estimated by a separate probabilistic model, so the estimation error can be inherited in the confidence information itself. In contrast, the toy experiment uses exact confidences from the data-generating model, where Pconf follows the predicted convergence trend.

We next evaluate whether weak labels are also useful for post-hoc correction. Figures 2 and 3 compare oracle MC on held-out test data before and after correction under $(\mathrm{PN}, \mathrm{PN})$, $(\mathrm{PN}, \mathrm{weak})$, and $(\mathrm{weak}, \mathrm{weak})$. Each panel pools experimental runs: $240 = 10$ seeds $\times\, 4$ datasets $\times\, 6$ base predictors for $(\mathrm{PN}, \mathrm{PN})$; the $(\mathrm{PN}, \mathrm{weak})$ and $(\mathrm{weak}, \mathrm{weak})$ panels have 720 runs each by additionally pooling PU, UU, and Pconf. Points below the diagonal indicate improvement, and the number in each panel reports how many runs reduce oracle MC. Figure 2 shows results for our proposed WLMC, whereas Figure 3 shows results for the method selected from WLMC, temperature scaling, and Platt scaling using the smallest oracle validation MC. In $(\mathrm{PN}, \mathrm{PN})$, WLMC uses the boosting-style updates of Algorithm 1 with clean labels to estimate calibration violations. Figure 3 uses oracle MC for selection.

The left panels of Figures 2 and 3 show that correction in $(\mathrm{PN}, \mathrm{PN})$ does not always improve MC, even with label-based boosting updates or validation selection among the three methods. This is partly because PN-trained baselines often already have small violations. The middle panel of Figure 2 shows that WLMC can improve PN-trained predictors without additional clean labels. The right panel shows that WLMC can also reduce MC when both base learning and correction use weak data. Weakly trained predictors typically start with larger MC than PN-trained predictors, as also shown in Appendix K, giving post-processing more room to improve. Figure 3 shows the same overall

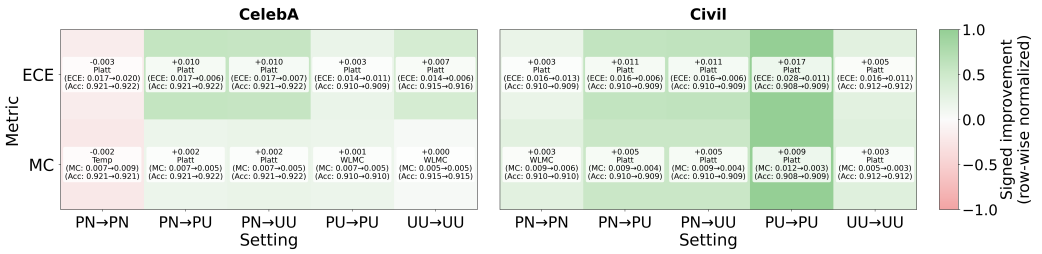


Figure 4: ECE and MC improvements on CelebA and CivilComments.

tendency under its selection protocol: weak data can support post-hoc calibration of both PN-trained and weakly trained predictors.

Image and text experiments with larger models. Finally, we test whether the same phenomena appear for larger neural models, using CelebA [30] with a ViT-based image classifier [9] and CivilComments [3] with a BERT-based text classifier [38]. For each setting, we freeze the baseline predictor and compare WLMC, temperature scaling, and Platt scaling. Experimental details are reported in Appendices I and L; Appendix L also reports a CelebA-ResNet [22] experiment.

Figure 4 reports, for each metric, the method with the largest improvement together with the change in accuracy. Weak-label post-processing remains effective beyond tabular data. ECE and MC improve even in the $(\mathrm{PN}, \mathrm{weak})$ setting, and the ECE improvement is sometimes larger than in $(\mathrm{PN}, \mathrm{PN})$, despite using weaker supervision for correction. The $(\mathrm{weak}, \mathrm{weak})$ settings also improve consistently, although their baselines tend to have larger ECE and MC compared with PN baselines.

Across methods, Platt scaling is strong, especially for ECE, while WLMC is competitive for MC and performs comparably to Platt scaling in several CelebA settings. This is consistent with clean-label reports that Platt scaling can be a strong calibration baseline; we do not aim to outperform Platt scaling in every case, but instead show that weak-label post-processing, including WLMC, can achieve comparable calibration improvements under weak supervision. Temperature scaling is less reliable in some CelebA configurations, suggesting that a global temperature may be insufficient for subgroup-level errors. Importantly, none of the displayed methods causes a large accuracy drop, indicating that weak-label post-processing can improve calibration without sacrificing predictive performance.

7 Conclusion and limitations

In this work, we developed a unified framework for multicalibration under weak supervision by expressing calibration witnesses through the decontamination principle used in WSL. This yields corrected estimators for PU, UU, and Pconf data, together with finite-sample guarantees, and enables post-hoc correction using weak labels rather than clean calibration labels. Empirically, to the best of our knowledge, this is the first numerical study of weak-label multicalibration on real-data benchmarks with controlled weak-observation views, covering estimation and post-hoc correction. The results show that weak estimates can track oracle multicalibration, weak training tends to increase calibration violations, and weak labels can still improve post-hoc calibration, suggesting reduced reliance on clean calibration labels. The main limitations are the need for reliable class priors and mixture proportions and the binary-label focus of our methods. Improving robustness to estimated nuisance parameters and extending the framework to multiclass settings are important directions for future work.

Acknowledgements

FF was supported by JST BOOST Program Grant Number JPMJBY24G8. This work was supported by Japan Science and Technology Agency (JST) as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE), Grant Number JPMJAP25B1.

References

[1] Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

[2] Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1727–1740, 2023.

[3] Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

[4] Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Predictive inference with weak supervision. *Journal of Machine Learning Research*, 25(118):1–45, 2024.

[5] Chiang, C.-K. and Sugiyama, M. Unified risk analysis for weakly supervised learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[6] Cooper, A. F., Barocas, S., De Sa, C., and Sen, S. Variance, self-consistency, and arbitrariness in fair classification. *arXiv preprint arXiv:2301.11562*, pp. 1–84, 2023.

[7] Dawid, A. P. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

[8] Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[11] Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1095–1108, 2021.

[12] Einbinder, B.-S., Feldman, S., Bates, S., Angelopoulos, A. N., Gendler, A., and Romano, Y. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25 (328):1–66, 2024.

[13] Foster, D. P. and Vohra, R. V. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[14] Futami, F. and Fujisawa, M. Information-theoretic generalization analysis for expected calibration error. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84246–84297. Curran Associates, Inc., 2024.

[15] Futami, F. and Nitanda, A. Smooth calibration error: Uniform convergence and functional gradient analysis. In *The Fourteenth International Conference on Learning Representations*, 2026.

[16] Globus-Harris, I., Harrison, D., Kearns, M., Roth, A., and Sorrell, J. Multicalibration as boosting for regression. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11459–11492. PMLR, 23–29 Jul 2023.

[17] Gopalan, P., Kim, M. P., Singhal, M. A., and Zhao, S. Low-degree multicalibration. In *Conference on Learning Theory*, pp. 3193–3234. PMLR, 2022.

[18] Gopalan, P., Hu, L., and Rothblum, G. N. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1983–2026. PMLR, 2024.

[19] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330, 2017.

[20] Gupta, C., Podkopaev, A., and Ramdas, A. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

[21] Hansen, D., Devic, S., Nakkiran, P., and Sharan, V. When is multicalibration post-processing necessary? *Advances in Neural Information Processing Systems*, 37:38383–38455, 2024.

[22] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[23] Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018.

[24] Hu, L., Jambulapati, A., Tian, K., and Yang, C. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[25] Ishida, T., Niu, G., and Sugiyama, M. Binary classification from positive-confidence data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[26] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018.

[27] Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

[28] Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[29] Kiryo, R., Futami, F., and Sugiyama, M. Estimating expected calibration error for positive-unlabeled learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

[30] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[31] Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. In *International Conference on Learning Representations*, 2019.

[32] Lu, N., Lei, S., Niu, G., Sato, I., and Sugiyama, M. Binary classification from multiple unlabeled datasets via surrogate set classification. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7134–7144. PMLR, 18–24 Jul 2021.

[33] Marx, C., Zalouk, S., and Ermon, S. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36:25910–25928, 2023.

[34] Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2 edition, 2018. ISBN 978-0-262-03940-6.

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

[36] Plessis, M. D., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1386–1394, Lille, France, 07–09 Jul 2015. PMLR.

[37] Rossellini, R., Soloff, J. A., Barber, R. F., Ren, Z., and Willett, R. Can a calibration metric be both testable and actionable? In Haghtalab, N. and Moitra, A. (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4937–4972. PMLR, 30 Jun–04 Jul 2025.

[38] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[39] Sharma, S., Gee, A. H., Paydarfar, D., and Ghosh, J. Fair-n: Fair and robust neural networks for structured data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 946–955, 2021.

[40] Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021.

[41] Yeh, I.-C. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

A Proofs for the generic rewrite statements

A.1 Review of the risk rewrite

Starting from Eq. (5), write

$$R_\ell(f) = \int_{\mathcal{X}} L_\ell(x; f)^\top \, dB(x).$$

Substituting the decontamination identity $B = M_{\mathrm{corr}}^\dagger \bar{P}$ from Eq. (6) gives

$$R_\ell(f) = \int_{\mathcal{X}} L_\ell(x; f)^\top M_{\mathrm{corr}}^\dagger \, d\bar{P}(x),$$

which is exactly Eq. (7).

The same substitution also covers the x-dependent case that appears in confidence-based weak supervision. Namely, whenever the clean conditional vector can be recovered through an x-dependent decontamination operator

$$B(dx) = M_{\mathrm{corr}}^\dagger(x) \, \bar{P}(dx),$$

we obtain the pointwise rewrite

$$R_\ell(f) = \int_{\mathcal{X}} L_\ell(x; f)^\top M_{\mathrm{corr}}^\dagger(x) \, d\bar{P}(x).$$

Thus the algebra of the contamination-matrix case remains unchanged except that the correction coefficients are now functions of x. In the terminology of Chiang & Sugiyama [5], this is the regime of confidence-based settings, where observable confidences provide the weights defining the contamination matrix. Section 4.2 treats Pconf as the representative binary example, and Appendix B records other instances such as posterior-confidence and Sconf supervision.

A.2 Proof of Theorem 1

By definition,

$$R_{c,w}(f) = \mathbb{E}[c(X)w(f(X))(Y - f(X))].$$

Fix (c, w). By conditioning on Y, we obtain

$$\begin{aligned} R_{c,w}(f) &= \mathbb{E}[\phi_f^{c,w}(X, Y)] \\ &= \mathbb{E}[\phi_f^{c,w}(X, Y)\mathbf{1}\{Y = 1\}] + \mathbb{E}[\phi_f^{c,w}(X, Y)\mathbf{1}\{Y = 0\}] \\ &= P(Y = 1)\mathbb{E}[\phi_f^{c,w}(X, 1) \mid Y = 1] + P(Y = 0)\mathbb{E}[\phi_f^{c,w}(X, 0) \mid Y = 0] \\ &= \pi_+ \mathbb{E}_{P_+}[\phi_f^{c,w}(X, 1)] + \pi_- \mathbb{E}_{P_-}[\phi_f^{c,w}(X, 0)]. \end{aligned}$$

For a fixed witness ϕ_f, define the corresponding vector of label-resolved terms

$$L_{c,w}(x; f) := \begin{pmatrix} \pi_+ \phi_f(x, 1) \\ \pi_- \phi_f(x, 0) \end{pmatrix}.$$

Consequently, the witness moment is

$$R_{c,w}(f) = \int_{\mathcal{X}} L_{c,w}(x; f)^\top \, dB(x).$$

Substituting the decontamination identity $B = M_{\mathrm{corr}}^\dagger \bar{P}$ from Eq. (6) gives

$$R_{c,w}(f) = \int_{\mathcal{X}} L_{c,w}(x; f)^\top M_{\mathrm{corr}}^\dagger \, d\bar{P}(x).$$

Taking the supremum over (c, w) yields Eq. (11).

B Additional weak-supervision settings

This appendix collects the binary weak-supervision estimators induced by the contamination-based rewrite. Following the request of the main text, we omit multiclass cases and organize the appendix by weak-supervision model. In every case, the derivation follows the same template: first rewrite a fixed witness

$$R_\phi(f) := \pi_+ \mathbb{E}_{P_+}[\phi_f(X,1)] + \pi_- \mathbb{E}_{P_-}[\phi_f(X,0)],$$

under the observable weak distributions, and then substitute the multicalibration witness

$$\phi_f^{c,w}(x,y) = c(x)w(f(x))(y - f(x)).$$

For later use, note that

$$\phi_f^{c,w}(x,1) = c(x)w(f(x))(1-f(x)), \qquad \phi_f^{c,w}(x,0) = -c(x)w(f(x))f(x).$$

B.1 MCD scenarios: UU template and special cases

We begin with the matrix-contamination settings. This subsection gives the derivation underlying the concise UU discussion in Section 4.2.

Data model. In the general UU setting we observe two unlabeled samples $X_1^{U_1}, \dots, X_{n_1}^{U_1} \sim P_{U_1}$ and $X_1^{U_2}, \dots, X_{n_2}^{U_2} \sim P_{U_2}$. Write $\theta_k = P(Y=1 \mid U_k)$, so that $P_{U_k} = \theta_k P_+ + (1-\theta_k)P_-$. We use $B = (P_+, P_-)^\top$ and $L_\phi(x) = (\pi_+\phi_f(x,1), \pi_-\phi_f(x,0))^\top$. Then

$$\bar{P} = \begin{pmatrix} P_{U_1} \\ P_{U_2} \end{pmatrix} = M_{\mathrm{UU}} B, \qquad M_{\mathrm{UU}} = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ \theta_2 & 1-\theta_2 \end{pmatrix},$$

with $\Delta := \theta_1 - \theta_2 \neq 0$. Hence

$$M_{\mathrm{UU}}^{-1} = \frac{1}{\Delta} \begin{pmatrix} 1-\theta_2 & -(1-\theta_1) \\ -\theta_2 & \theta_1 \end{pmatrix}.$$

Multiplying $L_\phi(x)^\top$ by M_{UU}^{-1} gives the corrected witness row vector

$$L_\phi(x)^\top M_{\mathrm{UU}}^{-1} = \frac{1}{\Delta}\Big((1-\theta_2)\pi_+\phi_f(x,1) - \theta_2\pi_-\phi_f(x,0), \\ -(1-\theta_1)\pi_+\phi_f(x,1) + \theta_1\pi_-\phi_f(x,0)\Big). \tag{15}$$

Substituting Eq. (15) into the generic rewrite yields

$$R_\phi^{\mathrm{UU}}(f) = \mathbb{E}_{P_{U_1}}\left[\frac{(1-\theta_2)\pi_+}{\Delta}\phi_f(X,1) - \frac{\theta_2\pi_-}{\Delta}\phi_f(X,0)\right] \\ + \mathbb{E}_{P_{U_2}}\left[-\frac{(1-\theta_1)\pi_+}{\Delta}\phi_f(X,1) + \frac{\theta_1\pi_-}{\Delta}\phi_f(X,0)\right]. \tag{16}$$

For the multicalibration witness, define

$$\alpha_1(v) := \frac{(1-\theta_2)\pi_+(1-v) + \theta_2\pi_- v}{\Delta}, \qquad \alpha_2(v) := \frac{-(1-\theta_1)\pi_+(1-v) - \theta_1\pi_- v}{\Delta}.$$

With this notation,

$$R_{c,w}^{\mathrm{UU}}(f) = \sum_{k=1}^{2} \mathbb{E}_{P_{U_k}}[\alpha_k(f(X))c(X)w(f(X))], \\ \widehat{R}_{c,w}^{\mathrm{UU}}(f) := \sum_{k=1}^{2} \frac{1}{n_k}\sum_{i=1}^{n_k} \alpha_k(f(X_i^{U_k}))c(X_i^{U_k})w(f(X_i^{U_k})). \tag{17}$$

The remaining MCD-type estimators are obtained by substituting the appropriate positive-class proportions into Eq. (16)–(17). Unless stated otherwise, these MCD-style rewrites assume that the class prior π_+ and the weak-source mixing parameters appearing in the model are known or have been pre-estimated. We write the resulting fixed-witness identities explicitly below. In the similar/dissimilar settings, $P_{\tilde{S}}$ and $P_{\tilde{D}}$ denote the observable one-point marginals induced by similar and dissimilar pair supervision, namely

$$P_{\tilde{S}} = \frac{\pi_+^2}{\pi_+^2+\pi_-^2}P_+ + \frac{\pi_-^2}{\pi_+^2+\pi_-^2}P_-, \qquad P_{\tilde{D}} = \frac{1}{2}P_+ + \frac{1}{2}P_-.$$

PU learning. *Observed data.* We observe a positively labeled sample $X_1^+, \dots, X_{n_+}^+ \sim P_+$ and an unlabeled sample $X_1^u, \dots, X_{n_u}^u \sim P_X = \pi_+ P_+ + \pi_- P_-$. We assume the class prior π_+ is known or pre-estimated. This is the UU specialization $(\theta_1, \theta_2) = (1, \pi_+)$. Set $\theta_1 = 1$ and $\theta_2 = \pi_+$, so that

$$\bar{P} = \begin{pmatrix} P_+ \\ P_X \end{pmatrix}, \qquad M_{\mathrm{PU}} = \begin{pmatrix} 1 & 0 \\ \pi_+ & \pi_- \end{pmatrix}, \qquad M_{\mathrm{PU}}^{-1} = \begin{pmatrix} 1 & 0 \\ -\pi_+/\pi_- & 1/\pi_- \end{pmatrix}.$$

Hence the PU fixed-witness rewrite is

$$R_\phi^{\mathrm{PU}}(f) = \pi_+ \mathbb{E}_{P_+}[\phi_f(X,1) - \phi_f(X,0)] + \mathbb{E}_{P_X}[\phi_f(X,0)].$$

Since $\phi_f^{c,w}(x,1) - \phi_f^{c,w}(x,0) = c(x)w(f(x))$, it follows that

$$\begin{aligned} R_{c,w}^{\mathrm{PU}}(f) &= \pi_+ \mathbb{E}_{P_+}[c(X)w(f(X))] - \mathbb{E}_{P_X}[c(X)w(f(X))f(X)], \\ \widehat{R}_{c,w}^{\mathrm{PU}}(f) &:= \pi_+ \frac{1}{n_+} \sum_{i=1}^{n_+} c(X_i^+)w(f(X_i^+)) - \frac{1}{n_u} \sum_{j=1}^{n_u} c(X_j^u)w(f(X_j^u))f(X_j^u). \end{aligned} \tag{18}$$

SU learning. *Observed data.* We observe i.i.d. similar pairs $(X_i^s, X_i^{s\prime})$ known to share the same label, together with unlabeled points $X_1^u, \dots, X_{n_u}^u \sim P_X$. Let $X_1^{\tilde{S}}, \dots, X_{n_s}^{\tilde{S}} \sim P_{\tilde{S}}$ denote the induced one-point marginal of the similar-pair sample. The rewrite below is written in terms of this observable marginal. Set $\theta_1 = \pi_+^2/(\pi_+^2 + \pi_-^2)$ and $\theta_2 = \pi_+$. Then

$$\begin{aligned} R_\phi^{\mathrm{SU}}(f) = \mathbb{E}_{P_{\tilde{S}}}\left[\frac{\pi_+^2 + \pi_-^2}{\pi_+ - \pi_-}\big(\phi_f(X,1) - \phi_f(X,0)\big)\right] \\ + \mathbb{E}_{P_X}\left[-\frac{\pi_-}{\pi_+ - \pi_-}\phi_f(X,1) + \frac{\pi_+}{\pi_+ - \pi_-}\phi_f(X,0)\right]. \end{aligned}$$

Substituting $\phi_f^{c,w}$ gives

$$\begin{aligned} R_{c,w}^{\mathrm{SU}}(f) &= \frac{\pi_+^2 + \pi_-^2}{\pi_+ - \pi_-}\mathbb{E}_{P_{\tilde{S}}}[c(X)w(f(X))] - \mathbb{E}_{P_X}\left[c(X)w(f(X))\left(\frac{\pi_-}{\pi_+ - \pi_-} + f(X)\right)\right], \\ \widehat{R}_{c,w}^{\mathrm{SU}}(f) &:= \frac{\pi_+^2 + \pi_-^2}{\pi_+ - \pi_-}\frac{1}{n_s}\sum_{i=1}^{n_s} c(X_i^{\tilde{S}})w(f(X_i^{\tilde{S}})) \\ &\quad - \frac{1}{n_u}\sum_{j=1}^{n_u} c(X_j^u)w(f(X_j^u))\left(\frac{\pi_-}{\pi_+ - \pi_-} + f(X_j^u)\right). \end{aligned} \tag{19}$$

DU learning. *Observed data.* We observe i.i.d. dissimilar pairs $(X_i^d, X_i^{d\prime})$ known to have different labels, together with unlabeled points $X_1^u, \dots, X_{n_u}^u \sim P_X$. Let $X_1^{\tilde{D}}, \dots, X_{n_d}^{\tilde{D}} \sim P_{\tilde{D}}$ denote the induced one-point marginal of the dissimilar-pair sample. The rewrite below is written in terms of this observable marginal. Set $\theta_1 = 1/2$ and $\theta_2 = \pi_+$. Then

$$\begin{aligned} R_\phi^{\mathrm{DU}}(f) = \mathbb{E}_{P_{\tilde{D}}}\left[\frac{2\pi_+\pi_-}{\pi_- - \pi_+}\big(\phi_f(X,1) - \phi_f(X,0)\big)\right] \\ + \mathbb{E}_{P_X}\left[-\frac{\pi_+}{\pi_- - \pi_+}\phi_f(X,1) + \frac{\pi_-}{\pi_- - \pi_+}\phi_f(X,0)\right]. \end{aligned}$$

Therefore the DU witness moment is

$$\begin{aligned} R_{c,w}^{\mathrm{DU}}(f) &= -\frac{2\pi_+\pi_-}{\pi_+ - \pi_-}\mathbb{E}_{P_{\tilde{D}}}[c(X)w(f(X))] + \mathbb{E}_{P_X}\left[c(X)w(f(X))\left(\frac{\pi_+}{\pi_+ - \pi_-} - f(X)\right)\right], \\ \widehat{R}_{c,w}^{\mathrm{DU}}(f) &:= -\frac{2\pi_+\pi_-}{\pi_+ - \pi_-}\frac{1}{n_d}\sum_{i=1}^{n_d} c(X_i^{\tilde{D}})w(f(X_i^{\tilde{D}})) \\ &\quad + \frac{1}{n_u}\sum_{j=1}^{n_u} c(X_j^u)w(f(X_j^u))\left(\frac{\pi_+}{\pi_+ - \pi_-} - f(X_j^u)\right). \end{aligned} \tag{20}$$

SD learning. *Observed data.* We observe both i.i.d. similar pairs $(X_i^s, X_i^{s\prime})$ and i.i.d. dissimilar pairs $(X_j^d, X_j^{d\prime})$. Let $X_1^{\tilde{S}}, \ldots, X_{n_s}^{\tilde{S}} \sim P_{\tilde{S}}$ and $X_1^{\tilde{D}}, \ldots, X_{n_d}^{\tilde{D}} \sim P_{\tilde{D}}$ denote the corresponding one-point marginals used below. Set $\theta_1 = \pi_+^2/(\pi_+^2 + \pi_-^2)$ and $\theta_2 = 1/2$. Then

$$R_\phi^{\mathrm{SD}}(f) = (\pi_+^2 + \pi_-^2)\mathbb{E}_{P_{\tilde{S}}}\left[\frac{\pi_+}{\pi_+ - \pi_-}\phi_f(X,1) - \frac{\pi_-}{\pi_+ - \pi_-}\phi_f(X,0)\right] + 2\pi_+\pi_-\mathbb{E}_{P_{\tilde{D}}}\left[-\frac{\pi_-}{\pi_+ - \pi_-}\phi_f(X,1) + \frac{\pi_+}{\pi_+ - \pi_-}\phi_f(X,0)\right].$$

Therefore the SD witness moment is

$$\begin{aligned} R_{c,w}^{\mathrm{SD}}(f) &= (\pi_+^2 + \pi_-^2)\mathbb{E}_{P_{\tilde{S}}}\left[c(X)w(f(X))\left(\frac{\pi_+}{\pi_+ - \pi_-} - f(X)\right)\right] \\ &\quad - 2\pi_+\pi_-\mathbb{E}_{P_{\tilde{D}}}\left[c(X)w(f(X))\left(\frac{\pi_-}{\pi_+ - \pi_-} + f(X)\right)\right], \\ \widehat{R}_{c,w}^{\mathrm{SD}}(f) &:= (\pi_+^2 + \pi_-^2)\frac{1}{n_s}\sum_{i=1}^{n_s} c(X_i^{\tilde{S}})w(f(X_i^{\tilde{S}}))\left(\frac{\pi_+}{\pi_+ - \pi_-} - f(X_i^{\tilde{S}})\right) \\ &\quad - 2\pi_+\pi_-\frac{1}{n_d}\sum_{j=1}^{n_d} c(X_j^{\tilde{D}})w(f(X_j^{\tilde{D}}))\left(\frac{\pi_-}{\pi_+ - \pi_-} + f(X_j^{\tilde{D}})\right). \end{aligned} \tag{21}$$

B.2 Pairwise comparison learning

Observed data. We observe comparison pairs carrying a superior/inferior outcome rather than a clean binary label. Let $X_1^{\mathrm{sup}}, \ldots, X_{n_{\mathrm{sup}}}^{\mathrm{sup}} \sim P_{\mathrm{Sup}}$ and $X_1^{\mathrm{inf}}, \ldots, X_{n_{\mathrm{inf}}}^{\mathrm{inf}} \sim P_{\mathrm{Inf}}$ denote the induced one-point marginals associated with superior and inferior events, respectively. The comparison model has marginals

$$P_{\mathrm{Sup}} = \frac{\pi_+ P_+ + \pi_-^2 P_-}{1 - \pi_+\pi_-}, \qquad P_{\mathrm{Inf}} = \frac{\pi_+^2 P_+ + \pi_- P_-}{1 - \pi_+\pi_-}.$$

The fixed-witness rewrite uses these marginals and the known class prior. The unified-risk rewrite is expressed through the corrected losses

$$\bar{\ell}_{\mathrm{Sup}} = \ell_1 - \pi_+\ell_0, \qquad \bar{\ell}_{\mathrm{Inf}} = -\pi_-\ell_1 + \ell_0,$$

where $\mathbb{E}_{\mathrm{Sup}}$ and $\mathbb{E}_{\mathrm{Inf}}$ denote expectations over P_{Sup} and P_{Inf}. Replacing (ℓ_1, ℓ_0) by $(\phi_f(\cdot,1), \phi_f(\cdot,0))$ yields

$$R_\phi^{\mathrm{Pcomp}}(f) = \mathbb{E}_{\mathrm{Sup}}[\phi_f(X,1) - \pi_+\phi_f(X,0)] + \mathbb{E}_{\mathrm{Inf}}[-\pi_-\phi_f(X,1) + \phi_f(X,0)].$$

Thus the comparison-learning witness moment is

$$\begin{aligned} R_{c,w}^{\mathrm{Pcomp}}(f) &= \mathbb{E}_{\mathrm{Sup}}[c(X)w(f(X))(1 - \pi_- f(X))] - \mathbb{E}_{\mathrm{Inf}}[c(X)w(f(X))(\pi_- + \pi_+ f(X))], \\ \widehat{R}_{c,w}^{\mathrm{Pcomp}}(f) &:= \frac{1}{n_{\mathrm{sup}}}\sum_{i=1}^{n_{\mathrm{sup}}} c(X_i^{\mathrm{sup}})w(f(X_i^{\mathrm{sup}}))(1 - \pi_- f(X_i^{\mathrm{sup}})) \\ &\quad - \frac{1}{n_{\mathrm{inf}}}\sum_{j=1}^{n_{\mathrm{inf}}} c(X_j^{\mathrm{inf}})w(f(X_j^{\mathrm{inf}}))(\pi_- + \pi_+ f(X_j^{\mathrm{inf}})). \end{aligned} \tag{22}$$

B.3 Confidence-based binary settings

We next collect binary settings whose decontamination rule depends on observed confidences.

Posterior-confidence learning. *Observed data.* We observe i.i.d. pairs (X_i, r_i) with $X_i \sim P_X$ and $r_i = r(X_i) = P(Y = 1 \mid X_i)$. In contrast to Pconf, the sample is drawn from the population marginal rather than from P_+. Conditioning on X directly gives

$$R_\phi^{\mathrm{PC}}(f) = \mathbb{E}_X\big[r(X)\phi_f(X,1) + (1 - r(X))\phi_f(X,0)\big].$$

Specializing to $\phi_f^{c,w}$ yields

$$R_{c,w}^{\mathrm{PC}}(f) = \mathbb{E}_X\big[c(X)w(f(X))(r(X) - f(X))\big],$$
$$\widehat{R}_{c,w}^{\mathrm{PC}}(f) := \frac{1}{n}\sum_{i=1}^{n} c(X_i)w(f(X_i))(r_i - f(X_i)). \tag{23}$$

Thus this supervision model replaces the inaccessible Bernoulli outcome Y by the observed posterior mean $r(X)$.

Pconf learning. *Observed data.* We observe i.i.d. positive-confidence pairs (X_i, r_i) with $X_i \sim P_+$ and $r_i = r(X_i) = P(Y = 1 \mid X_i)$. We also assume the class prior π_+ is known or pre-estimated, because the exact witness moment keeps the original population scaling. The exact rewrite requires $P_- \ll P_+$. In the $B = (P_+, P_-)^\top$ notation of Section 3.2, this corresponds to the x-dependent decontamination operator $M_{\mathrm{Pconf}}^{\dagger}(x) = \mathrm{diag}\{1, (\pi_+/\pi_-)(1 - r(x))/r(x)\}$ acting on the duplicated positive source $\bar{P}_{\mathrm{Pconf}} = (P_+, P_+)^\top$. Bayes' rule implies the density-ratio identity

$$\pi_- P_-(dx) = \pi_+ \frac{1 - r(x)}{r(x)} P_+(dx). \tag{24}$$

Substituting Eq. (24) into the label-resolved expansion of $R_\phi(f)$ gives

$$R_\phi^{\mathrm{Pconf}}(f) = \pi_+ \mathbb{E}_{P_+}\left[\phi_f(X, 1) + \frac{1 - r(X)}{r(X)}\phi_f(X, 0)\right].$$

For the multicalibration witness,

$$R_{c,w}^{\mathrm{Pconf}}(f) = \pi_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{r(X)}\right)\right],$$
$$\widehat{R}_{c,w}^{\mathrm{Pconf}}(f) := \frac{\pi_+}{n}\sum_{i=1}^{n} c(X_i)w(f(X_i))\left(1 - \frac{f(X_i)}{r_i}\right). \tag{25}$$

This is the exact fixed-witness estimator used in the main text.

Sconf learning. *Observed data.* We observe i.i.d. triples

$$(X_1, X_1', r_1), \ldots, (X_{n_{\mathrm{sc}}}, X_{n_{\mathrm{sc}}}', r_{n_{\mathrm{sc}}}),$$
$$(X_i, X_i') \sim P_{XX'}, \qquad r_i = r(X_i, X_i') := P(Y_i = Y_i' \mid X_i, X_i').$$

Each pair consists of independent draws $(X, Y), (X', Y')$ from P, so $P_{XX'} = P_X \otimes P_X$. The class prior π_+ is known and $\pi_+ \neq \pi_-$. In contrast to Pconf, the side information is attached to a pair rather than to a single positive example. The unified-risk rewrite gives

$$R_\phi^{\mathrm{Sconf}}(f) = \mathbb{E}_{(X,X',r)}\left[\frac{r - \pi_-}{\pi_+ - \pi_-}\frac{\phi_f(X, 1) + \phi_f(X', 1)}{2} + \frac{\pi_+ - r}{\pi_+ - \pi_-}\frac{\phi_f(X, 0) + \phi_f(X', 0)}{2}\right].$$

Substituting $\phi_f^{c,w}$ yields

$$R_{c,w}^{\mathrm{Sconf}}(f) = \mathbb{E}_{(X,X',r)}\left[\frac{r - \pi_-}{\pi_+ - \pi_-}\frac{c(X)w(f(X))(1 - f(X)) + c(X')w(f(X'))(1 - f(X'))}{2}\right.$$
$$\left. - \frac{\pi_+ - r}{\pi_+ - \pi_-}\frac{c(X)w(f(X))f(X) + c(X')w(f(X'))f(X')}{2}\right],$$
$$\widehat{R}_{c,w}^{\mathrm{Sconf}}(f) := \frac{1}{n_{\mathrm{sc}}}\sum_{i=1}^{n_{\mathrm{sc}}}\left[\frac{r_i - \pi_-}{\pi_+ - \pi_-}\frac{c(X_i)w(f(X_i))(1 - f(X_i)) + c(X_i')w(f(X_i'))(1 - f(X_i'))}{2}\right.$$
$$\left. - \frac{\pi_+ - r_i}{\pi_+ - \pi_-}\frac{c(X_i)w(f(X_i))f(X_i) + c(X_i')w(f(X_i'))f(X_i')}{2}\right]. \tag{26}$$

Summary. Every binary weak-supervision model admitting the stated decontamination identity gives a fixed-witness estimator by the same substitution principle: rewrite $R_\phi(f)$ using the decontamination rule of the model, then insert the witness $\phi_f^{c,w}$. The resulting empirical multicalibration criterion is always

$$\widehat{\mathrm{MC}}^{\mathrm{WSL}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left|\widehat{R}^{\mathrm{WSL}}_{c,w}(f)\right|,$$

with $\widehat{R}^{\mathrm{WSL}}_{c,w}(f)$ chosen from Eq. (17), Eq. (18), (19), (20), (21), (22), (23), (25), or (26) according to the observation model.

C Proofs of the uniform convergence

Throughout this appendix, $f, \mathcal{C}, \mathcal{W}$ are fixed independently of the weak samples, $0 < \delta < 1$, and the correction uses the exact weak-observation parameters. We write

$$\mathfrak{R}_n(\mathcal{C}; P) := \mathbb{E}_{S\sim P^n,\,\sigma}\left[\sup_{c\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i c(X_i)\right]$$

for the expected Rademacher complexity of $\mathcal{C}$ under P, where $\sigma_i \in \{-1, +1\}$ are independent Rademacher variables.

C.1 Uniform convergence for UU

For the UU estimator in Eq. (14), define

$$\mathrm{MC}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |R^{\mathrm{UU}}_{c,w}(f)|, \qquad \widehat{\mathrm{MC}}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |\widehat{R}^{\mathrm{UU}}_{c,w}(f)|.$$

Define the source envelopes

$$A_1 := \sup_{v\in[0,1]} |\alpha_1(v)|, \qquad A_2 := \sup_{v\in[0,1]} |\alpha_2(v)|.$$

Proposition 2 (Uniform convergence for the UU weak estimate)**.** *Assume $|c(x)| \le 1$ for all $c \in \mathcal{C}$ and $|w(v)| \le 1$ for all $w \in \mathcal{W}$. Then for any $\eta \in (0, 1]$, with probability at least $1 - \delta$,*

$$\left|\widehat{\mathrm{MC}}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f)\right| \le 2\sum_{k=1}^{2} A_k \left(\mathfrak{R}_{n_k}(\mathcal{C}; P_{U_k}) + \eta + \sqrt{\frac{2\log N_\infty(\eta, \mathcal{W}) + \log(4/\delta)}{n_k}}\right).$$

Proof of Proposition 2. Fix $\eta \in (0, 1]$, and let $\widetilde{\mathcal{W}} \subseteq \mathcal{W}$ be a minimal η-net of $\mathcal{W}$ under $\|\cdot\|_\infty$, so that

$$|\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W}).$$

For $k \in \{1, 2\}$ and $w \in \widetilde{\mathcal{W}}$, define the source-specific classes

$$\mathcal{G}_{k,w}(f) := \{x \mapsto \alpha_k(f(x))w(f(x))c(x) : c \in \mathcal{C}\}, \qquad \mathcal{G}_k(f) := \bigcup_{w\in\widetilde{\mathcal{W}}} \mathcal{G}_{k,w}(f).$$

By construction, every function in $\mathcal{G}_k(f)$ is bounded in absolute value by A_k.

By symmetrization and bounded-differences concentration, with probability at least $1 - \delta$, simultaneously for $k = 1, 2$,

$$\sup_{g\in\mathcal{G}_k(f)} \left|\frac{1}{n_k}\sum_{i=1}^{n_k} g(X_i^{U_k}) - \mathbb{E}_{P_{U_k}}[g(X)]\right| \le 2\mathfrak{R}_{n_k}(\mathcal{G}_k(f); P_{U_k}) + A_k\sqrt{\frac{2\log(4/\delta)}{n_k}}.$$

We next bound the Rademacher complexity of $\mathcal{G}_k(f)$ by the complexity of $\mathcal{C}$ and the size of the weight net. For a fixed $w \in \widetilde{\mathcal{W}}$ and a fixed sample $x_1, \ldots, x_{n_k}$, the maps

$$\psi_i(t) := \alpha_k(f(x_i))w(f(x_i))t$$

satisfy $\psi_i(0) = 0$ and are A_k-Lipschitz. Therefore the contraction inequality gives

$$\widehat{\mathfrak{R}}_{n_k}(\mathcal{G}_{k,w}(f)) \le A_k \widehat{\mathfrak{R}}_{n_k}(\mathcal{C}).$$

We use the following finite-union consequence of the sub-Gaussian maximal inequality. Conditional on a sample $x_1, \dots, x_m$, let $\mathcal{F} = \bigcup_{\ell=1}^{M} \mathcal{F}_\ell$ and suppose that $|h(x_i)| \le \lambda$ for all $h \in \mathcal{F}$ and all i. Define

$$Z_\ell := \sup_{h \in \mathcal{F}_\ell} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i), \qquad \ell = 1, \dots, M.$$

Then $Z_\ell - \mathbb{E}_\sigma Z_\ell$ is $\lambda/\sqrt{m}$-sub-Gaussian by the bounded-difference inequality. Therefore, the sub-Gaussian maximal inequality gives

$$\widehat{\mathfrak{R}}_m(\mathcal{F}) = \mathbb{E}_\sigma \max_{\ell \le M} Z_\ell \le \max_{\ell \le M} \widehat{\mathfrak{R}}_m(\mathcal{F}_\ell) + \lambda \sqrt{\frac{2 \log M}{m}}.$$

Applying this with

$$\mathcal{F} = \mathcal{G}_k(f), \qquad \mathcal{F}_\ell = \mathcal{G}_{k,w_\ell}(f), \qquad M = |\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W}), \qquad \lambda = A_k,$$

and then taking expectations over the U_k-sample, yields

$$\mathfrak{R}_{n_k}(\mathcal{G}_k(f); P_{U_k}) \le A_k \mathfrak{R}_{n_k}(\mathcal{C}; P_{U_k}) + A_k \sqrt{\frac{2 \log N_\infty(\eta, \mathcal{W})}{n_k}}.$$

For the explicit constants, apply symmetrization and bounded-differences concentration to each $\mathcal{G}_{k,w}(f)$, and take a union bound over $k \in \{1, 2\}$ and $w \in \widetilde{\mathcal{W}}$. The contraction bound gives, with probability at least $1 - \delta$,

$$\sup_{\substack{c \in \mathcal{C} \\ w \in \widetilde{\mathcal{W}}}} \left| \widehat{R}^{\mathrm{UU}}_{c,w}(f) - R^{\mathrm{UU}}_{c,w}(f) \right| \le 2 \sum_{k=1}^{2} A_k \left(\mathfrak{R}_{n_k}(\mathcal{C}; P_{U_k}) + \sqrt{\frac{2 \log N_\infty(\eta, \mathcal{W}) + \log(4/\delta)}{n_k}} \right) \quad (27)$$

It remains to pass from the net $\widetilde{\mathcal{W}}$ to the full class $\mathcal{W}$. For any $w \in \mathcal{W}$, choose $\widetilde{w} \in \widetilde{\mathcal{W}}$ such that $\|w - \widetilde{w}\|_\infty \le \eta$. Then, for each $k \in \{1, 2\}$, uniformly over $c \in \mathcal{C}$,

$$\left| \mathbb{E}_{P_{U_k}} \left[\alpha_k(f(X)) c(X) \{ w(f(X)) - \widetilde{w}(f(X)) \} \right] \right| \le A_k \eta,$$

and the same bound holds for the empirical average,

$$\left| \frac{1}{n_k} \sum_{i=1}^{n_k} \alpha_k(f(X_i^{U_k})) c(X_i^{U_k}) \{ w(f(X_i^{U_k})) - \widetilde{w}(f(X_i^{U_k})) \} \right| \le A_k \eta.$$

Thus the approximation error from replacing w by $\widetilde{w}$ is at most $2A_k\eta$ for source k. Therefore,

$$\begin{aligned} &\sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left| \widehat{R}^{\mathrm{UU}}_{c,w}(f) - R^{\mathrm{UU}}_{c,w}(f) \right| \\ &\le 2 \sum_{k=1}^{2} A_k \left(\mathfrak{R}_{n_k}(\mathcal{C}; P_{U_k}) + \eta + \sqrt{\frac{2 \log N_\infty(\eta, \mathcal{W}) + \log(4/\delta)}{n_k}} \right). \end{aligned}$$

Finally,

$$\left| \widehat{\mathrm{MC}}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}^{\mathrm{UU}}_{\mathcal{C},\mathcal{W}}(f) \right| \le \sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left| \widehat{R}^{\mathrm{UU}}_{c,w}(f) - R^{\mathrm{UU}}_{c,w}(f) \right|,$$

which completes the proof. □

C.2 Uniform convergence for PU

This section proves Proposition 1. The proof is written directly for the two PU sources, rather than by introducing the composed class $x \mapsto c(x)w(f(x))$.

Proof of Proposition 1. Fix $\eta \in (0, 1]$, and let $\widetilde{\mathcal{W}} \subseteq \mathcal{W}$ be a minimal η-net of $\mathcal{W}$ under $\|\cdot\|_\infty$, so that

$$|\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W}).$$

For $w \in \widetilde{\mathcal{W}}$, define the source-specific classes

$$\mathcal{G}_{+,w}(f) := \{x \mapsto \pi_+ \, w(f(x))c(x) : c \in \mathcal{C}\}, \qquad \mathcal{G}_{u,w}(f) := \{x \mapsto f(x)w(f(x))c(x) : c \in \mathcal{C}\},$$

and set

$$\mathcal{G}_+(f) := \bigcup_{w \in \widetilde{\mathcal{W}}} \mathcal{G}_{+,w}(f), \qquad \mathcal{G}_u(f) := \bigcup_{w \in \widetilde{\mathcal{W}}} \mathcal{G}_{u,w}(f).$$

Every function in $\mathcal{G}_+(f)$ is bounded by π_+ in absolute value, and every function in $\mathcal{G}_u(f)$ is bounded by 1 in absolute value.

By symmetrization and bounded-differences concentration, with probability at least $1 - \delta$, simultaneously for the positive and unlabeled samples,

$$\sup_{g \in \mathcal{G}_+(f)} \left| \frac{1}{n_+} \sum_{i=1}^{n_+} g(X_i^+) - \mathbb{E}_{P_+}[g(X)] \right| \le 2\mathfrak{R}_{n_+}(\mathcal{G}_+(f); P_+) + \pi_+ \sqrt{\frac{2\log(4/\delta)}{n_+}},$$

and, for the unlabeled source,

$$\sup_{g \in \mathcal{G}_u(f)} \left| \frac{1}{n_u} \sum_{j=1}^{n_u} g(X_j^u) - \mathbb{E}_{P_X}[g(X)] \right| \le 2\mathfrak{R}_{n_u}(\mathcal{G}_u(f); P_X) + \sqrt{\frac{2\log(4/\delta)}{n_u}}.$$

We next bound the two Rademacher complexities by the complexity of $\mathcal{C}$ and the size of the weight net. For a fixed $w \in \widetilde{\mathcal{W}}$ and a fixed positive sample $x_1, \ldots, x_{n_+}$, the maps

$$\psi_i(t) := \pi_+ \, w(f(x_i)) \, t$$

satisfy $\psi_i(0) = 0$ and are π_+-Lipschitz. Therefore the contraction inequality gives

$$\widehat{\mathfrak{R}}_{n_+}(\mathcal{G}_{+,w}(f)) \le \pi_+ \, \widehat{\mathfrak{R}}_{n_+}(\mathcal{C}).$$

We use the following finite-union consequence of the sub-Gaussian maximal inequality. Conditional on a sample $x_1, \ldots, x_m$, let $\mathcal{F} = \bigcup_{\ell=1}^M \mathcal{F}_\ell$ and suppose that $|h(x_i)| \le \lambda$ for all $h \in \mathcal{F}$ and all i. Define

$$Z_\ell := \sup_{h \in \mathcal{F}_\ell} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i), \qquad \ell = 1, \ldots, M.$$

Then $Z_\ell - \mathbb{E}_\sigma Z_\ell$ is $\lambda/\sqrt{m}$-sub-Gaussian by the bounded-difference inequality. Therefore, the sub-Gaussian maximal inequality gives

$$\widehat{\mathfrak{R}}_m(\mathcal{F}) = \mathbb{E}_\sigma \max_{\ell \le M} Z_\ell \le \max_{\ell \le M} \widehat{\mathfrak{R}}_m(\mathcal{F}_\ell) + \lambda \sqrt{\frac{2\log M}{m}}.$$

Applying this with $M = |\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W})$ and $\lambda = \pi_+$, and then taking expectations over the positive sample, yields

$$\mathfrak{R}_{n_+}(\mathcal{G}_+(f); P_+) \le \pi_+ \, \mathfrak{R}_{n_+}(\mathcal{C}; P_+) + \pi_+ \sqrt{\frac{2\log N_\infty(\eta, \mathcal{W})}{n_+}}.$$

Similarly, for the unlabeled sample, the maps

$$\psi_i(t) := f(x_i)w(f(x_i)) \, t$$

satisfy $\psi_i(0) = 0$ and are 1-Lipschitz, because $0 \le f \le 1$ and $\|w\|_\infty \le 1$. Hence, by the contraction inequality and the same finite-union argument with $M = |\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W})$ and $\lambda = 1$,

$$\mathfrak{R}_{n_u}(\mathcal{G}_u(f); P_X) \le \mathfrak{R}_{n_u}(\mathcal{C}; P_X) + \sqrt{\frac{2 \log N_\infty(\eta, \mathcal{W})}{n_u}}.$$

The same argument as in Eq. (27), with $(P_{U_1}, n_1, A_1) = (P_+, n_+, \pi_+)$ and $(P_{U_2}, n_2, A_2) = (P_X, n_u, 1)$, gives the corresponding uniform bound for the PU estimator with probability at least $1 - \delta$.

It remains to pass from $\widetilde{\mathcal{W}}$ to $\mathcal{W}$. For any $w \in \mathcal{W}$, choose $\widetilde{w} \in \widetilde{\mathcal{W}}$ such that $\|w - \widetilde{w}\|_\infty \le \eta$. Then, uniformly over $c \in \mathcal{C}$,

$$\left|\pi_+ \, \mathbb{E}_{P_+} \left[c(X)\{w(f(X)) - \widetilde{w}(f(X))\}\right]\right| \le \pi_+ \eta,$$

and the same bound holds for the empirical positive average. Likewise,

$$\left|\mathbb{E}_{P_X} \left[c(X)\{w(f(X)) - \widetilde{w}(f(X))\} f(X)\right]\right| \le \eta,$$

and the same bound holds for the empirical unlabeled average. Absorbing these approximation errors yields the bound in Eq. (13). Finally,

$$\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f)\right| \le \sup_{c \in \mathcal{C},\, w \in \mathcal{W}} \left|\widehat{R}^{\mathrm{PU}}_{c,w}(f) - R^{\mathrm{PU}}_{c,w}(f)\right|,$$

which completes the proof. □

Example rates. The same bound yields explicit rates for other common witness choices. If $n = n_+ = n_u$, $\mathcal{W}$ is finite, and $\mathcal{C}$ has VC dimension d, the VC/Rademacher bound gives

$$\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}_{\mathcal{C},\mathcal{W}}(f)\right| = O_p\left(\sqrt{\frac{d + \log(2|\mathcal{W}|)}{n}}\right).$$

If $\mathcal{C} = \{1\}$ and $\mathcal{W} = \mathrm{Lip}_1([0,1], [-1,1])$, corresponding to SmoothCE, the preceding single-scale covering bound is not sharp. Indeed, since $\log N_\infty(\eta, \mathcal{W}) = O(1/\eta)$, optimizing Eq. (13) over one scale gives only $O_p(n^{-1/3})$. For SmoothCE we instead use Dudley's chaining bound for the one-dimensional Lipschitz witness class. Define

$$\mathcal{G}_1 := \{z \mapsto w(z) : w \in \mathcal{W}\}, \qquad \mathcal{G}_2 := \{z \mapsto z w(z) : w \in \mathcal{W}\}.$$

Then $\mathcal{G}_1 \subset \mathrm{Lip}_1([0,1], [-1,1])$. Moreover, for $w \in \mathcal{W}$, the map $z \mapsto z w(z)$ is bounded by one and is 2-Lipschitz, since

$$|zw(z) - z'w(z')| \le |z - z'|\,|w(z)| + z'|w(z) - w(z')| \le 2|z - z'|.$$

Hence, for $k = 1, 2$, there is a universal constant C such that

$$\log N_\infty(u, \mathcal{G}_k) \le \frac{C}{u}, \qquad u \in (0, 1].$$

For a positive sample $X_1^+, \dots, X_m^+ \overset{\text{i.i.d.}}{\sim} P_+$ and an unlabeled sample $X_1^u, \dots, X_m^u \overset{\text{i.i.d.}}{\sim} P_X$, write

$$\widehat{\mathbb{E}}_{P_+,m}[\varphi(X)] := \frac{1}{m} \sum_{i=1}^m \varphi(X_i^+), \qquad \widehat{\mathbb{E}}_{P_X,m}[\varphi(X)] := \frac{1}{m} \sum_{i=1}^m \varphi(X_i^u).$$

Dudley's entropy integral, with the empirical L_2 entropy bounded by the sup-norm entropy, gives, for each $k = 1, 2$,

$$\begin{aligned} &\mathbb{E}_{(X_i^+) \sim P_+^m} \sup_{g \in \mathcal{G}_k} \left|\widehat{\mathbb{E}}_{P_+,m}[g(f(X))] - \mathbb{E}_{P_+}[g(f(X))]\right| \\ &\le 2\mathfrak{R}_m(\mathcal{G}_k \circ f; P_+) \le \frac{C}{\sqrt{m}} \int_0^1 \sqrt{\log N_\infty(u, \mathcal{G}_k)}\, du \le \frac{C}{\sqrt{m}}, \end{aligned}$$

and similarly

$$\mathbb{E}_{(X_i^u)\sim P_X^m} \sup_{g\in\mathcal{G}_k} \left|\widehat{\mathbb{E}}_{P_X,m}[g(f(X))] - \mathbb{E}_{P_X}[g(f(X))]\right| \le 2\mathfrak{R}_m(\mathcal{G}_k \circ f; P_X) \le \frac{C}{\sqrt{m}}.$$

Here we used $\int_0^1 u^{-1/2}du < \infty$. Since all functions in $\mathcal{G}_k$ are uniformly bounded by one, bounded-difference concentration yields, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}_k} \left|\widehat{\mathbb{E}}_{P_+,m}[g(f(X))] - \mathbb{E}_{P_+}[g(f(X))]\right| \le C\sqrt{\frac{1+\log(1/\delta)}{m}},$$

and the same bound holds with P_+ and $\widehat{\mathbb{E}}_{P_+,m}$ replaced by P_X and $\widehat{\mathbb{E}}_{P_X,m}$. For the PU residual with $\mathcal{C} = \{1\}$,

$$\begin{aligned}
&\widehat{R}^{\mathrm{PU}}_{1,w}(f) - R^{\mathrm{PU}}_{1,w}(f) \\
&= \pi_+\Big\{\widehat{\mathbb{E}}_{P_+,n_+}[w(f(X))] - \mathbb{E}_{P_+}[w(f(X))]\Big\} \\
&\quad - \Big\{\widehat{\mathbb{E}}_{P_X,n_u}[f(X)w(f(X))] - \mathbb{E}_{P_X}[f(X)w(f(X))]\Big\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left|\widehat{R}^{\mathrm{PU}}_{1,w}(f) - R^{\mathrm{PU}}_{1,w}(f)\right| \\
&\le \pi_+ \sup_{g\in\mathcal{G}_1} \left|\widehat{\mathbb{E}}_{P_+,n_+}[g(f(X))] - \mathbb{E}_{P_+}[g(f(X))]\right| + \sup_{g\in\mathcal{G}_2} \left|\widehat{\mathbb{E}}_{P_X,n_u}[g(f(X))] - \mathbb{E}_{P_X}[g(f(X))]\right|.
\end{aligned}$$

Together with

$$\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\{1\},\mathcal{W}}(f) - \mathrm{MC}_{\{1\},\mathcal{W}}(f)\right| \le \sup_{w\in\mathcal{W}} \left|\widehat{R}^{\mathrm{PU}}_{1,w}(f) - R^{\mathrm{PU}}_{1,w}(f)\right|,$$

a union bound over the positive and unlabeled empirical processes gives, with probability at least $1-\delta$,

$$\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\{1\},\mathcal{W}}(f) - \mathrm{MC}_{\{1\},\mathcal{W}}(f)\right| \le C\left\{\pi_+\sqrt{\frac{1+\log(2/\delta)}{n_+}} + \sqrt{\frac{1+\log(2/\delta)}{n_u}}\right\}.$$

In particular, if $n_+ = n_u = n$, then

$$\left|\widehat{\mathrm{MC}}^{\mathrm{PU}}_{\{1\},\mathcal{W}}(f) - \mathrm{MC}_{\{1\},\mathcal{W}}(f)\right| = O_p(n^{-1/2}).$$

This result matches the order of smooth CE in Futami & Nitanda [15].

Remark 1 (VC-type dimension)**.** *We use the phrase "VC dimension" in the usual shorthand sense. When $\mathcal{C}$ is a class of indicator functions, this is the ordinary VC dimension. For a general $[0,1]$-valued subgroup class $\mathcal{C}$, the corresponding assumption should be interpreted as finite VC-subgraph dimension, equivalently finite pseudo-dimension up to standard convention-dependent constants. Under this condition the standard Rademacher bound gives*

$$\mathfrak{R}_n(\mathcal{C}; P) = O\left(\sqrt{\frac{d}{n}}\right),$$

which is the complexity term used in the example rates.

C.3 Uniform convergence for Pconf

This section states and proves the Pconf uniform convergence bound. The setup matches Ishida et al. [25]: the Pconf sample consists of i.i.d. draws (X_i, r_i) with $X_i \sim P_+$ and $r_i = r(X_i) = P(Y=1 \mid X_i)$.

Proposition 3 (Uniform convergence for the Pconf weak estimate)**.** *Define the population and empirical Pconf weak-estimate errors by*

$$\mathrm{MC}^{\mathrm{Pconf}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |R^{\mathrm{Pconf}}_{c,w}(f)|, \qquad \widehat{\mathrm{MC}}^{\mathrm{Pconf}}_{\mathcal{C},\mathcal{W}}(f) := \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |\widehat{R}^{\mathrm{Pconf}}_{c,w}(f)|.$$

Assume $f, \mathcal{C}, \mathcal{W}$ are fixed independently of the Pconf sample, $|c(x)| \le 1$ for all $c \in \mathcal{C}$, $|w(v)| \le 1$ for all $w \in \mathcal{W}$, and there exists $C_r > 0$ such that $r(x) \ge C_r$ almost surely under P_+. Then for any $\eta \in (0, 1]$, with probability at least $1 - \delta$,

$$\begin{aligned}&\left|\widehat{\mathrm{MC}}^{\mathrm{Pconf}}_{\mathcal{C},\mathcal{W}}(f) - \mathrm{MC}^{\mathrm{Pconf}}_{\mathcal{C},\mathcal{W}}(f)\right| \\ &\le C\,\pi_+\left(1 + \frac{1}{C_r}\right)\left(\mathfrak{R}_n(\mathcal{C}; P_+) + \eta + \sqrt{\frac{2\log N_\infty(\eta, \mathcal{W}) + \log(4/\delta)}{n}}\right).\end{aligned}$$

Proof. Fix $\eta \in (0, 1]$ and let $\widetilde{\mathcal{W}} \subseteq \mathcal{W}$ be a minimal η-net of $\mathcal{W}$ under $\|\cdot\|_\infty$, so $|\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W})$. Define

$$\mathcal{G}_w(f) := \left\{x \mapsto \pi_+\left(1 - \frac{f(x)}{r(x)}\right) w(f(x))c(x) : c \in \mathcal{C}\right\}, \qquad \mathcal{G}(f) := \bigcup_{w\in\widetilde{\mathcal{W}}} \mathcal{G}_w(f).$$

By Corollary 3,

$$\sup_{c\in\mathcal{C},\, w\in\widetilde{\mathcal{W}}} \left|\widehat{R}^{\mathrm{Pconf}}_{c,w}(f) - R^{\mathrm{Pconf}}_{c,w}(f)\right| = \sup_{g\in\mathcal{G}(f)} \left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_{P_+}[g(X)]\right|.$$

Because $0 \le f \le 1$, $|w| \le 1$, and $r(x) \ge C_r$ almost surely, every $g \in \mathcal{G}(f)$ satisfies

$$|g(x)| \le B := \pi_+\left(1 + \frac{1}{C_r}\right) \qquad P_+\text{-a.s.}$$

Standard symmetrization and McDiarmid's inequality therefore give, with probability at least $1 - \delta$,

$$\sup_{g\in\mathcal{G}(f)} \left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_{P_+}[g(X)]\right| \le 2\mathfrak{R}_n(\mathcal{G}(f); P_+) + B\sqrt{\frac{2\log(4/\delta)}{n}}.$$

For a fixed $w \in \widetilde{\mathcal{W}}$, the maps

$$\psi_i(t) := \pi_+\left(1 - \frac{f(x_i)}{r(x_i)}\right) w(f(x_i))t$$

satisfy $\psi_i(0) = 0$ and are B-Lipschitz, so the contraction inequality gives $\widehat{\mathfrak{R}}_n(\mathcal{G}_w(f)) \le B\widehat{\mathfrak{R}}_n(\mathcal{C})$. Taking expectations and applying the finite-union lemma over $|\widetilde{\mathcal{W}}| = N_\infty(\eta, \mathcal{W})$ classes yields

$$\mathfrak{R}_n(\mathcal{G}(f); P_+) \le B\mathfrak{R}_n(\mathcal{C}; P_+) + B\sqrt{\frac{2\log N_\infty(\eta, \mathcal{W})}{n}}.$$

It remains to move from the net back to the full class. For any $w \in \mathcal{W}$, choose $\tilde{w} \in \widetilde{\mathcal{W}}$ with $\|w - \tilde{w}\|_\infty \le \eta$. Then

$$\sup_c \left|\mathbb{E}_{P_+}\left[\pi_+\left(1 - \frac{f(X)}{r(X)}\right) c(X)(w(f(X)) - \tilde{w}(f(X)))\right]\right| \le B\eta,$$

and the same estimate holds for the empirical average. Absorbing these approximation errors proves the stated bound. □

Clipped Pconf estimator. The Pconf rewrite in Corollary 3 is exact but can have large variance when some confidences are small. A standard stabilization is to choose a clipping threshold $\tau \in (0, 1]$ and define

$$\widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f) := \frac{\pi_+}{n}\sum_{i=1}^n c(X_i)w(f(X_i))\left(1 - \frac{f(X_i)}{\tau \vee r_i}\right).$$

Inside Algorithm 1, one may replace $\widehat{R}^{\mathrm{Pconf}}$ by $\widehat{R}^{\mathrm{Pconf},\tau}$. Unless clipping leaves the confidences unchanged, the resulting bias must also be controlled.

D Analysis of nuisance parameters

The main paper assumes that the nuisance parameters used in the weak-supervision corrections are known. Here we analyze how errors in their estimates affect multicalibration estimation in the PU, UU, and Pconf settings. Assume $0 < \pi_+ < 1$, fix a predictor $f : \mathcal{X} \to [0,1]$, and let $\mathcal{C}$ and $\mathcal{W}$ be nonempty.

D.1 PU: estimated class prior

Given an estimate $\widehat{\pi}_+ \in [0,1]$ of the class prior π_+, define

$$e_\pi := |\widehat{\pi}_+ - \pi_+|, \qquad \widehat{\pi}_- := 1 - \widehat{\pi}_+.$$

The population residual using this estimate is

$$R^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+) := \widehat{\pi}_+ \mathbb{E}_{P_+}[c(X)w(f(X))] - \mathbb{E}_{P_X}[c(X)w(f(X))f(X)].$$

Proposition 4 (PU nuisance error)**.** *For every $(c,w) \in \mathcal{C} \times \mathcal{W}$,*

$$R^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+) - R_{c,w}(f) = (\widehat{\pi}_+ - \pi_+)\mathbb{E}_{P_+}[c(X)w(f(X))].$$

Consequently,

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left| R^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+) - R_{c,w}(f) \right| \le e_\pi. \tag{28}$$

Proof. Since $P_X = \pi_+ P_+ + \pi_- P_-$,

$$\begin{aligned} R_{c,w}(f) &= \pi_+ \mathbb{E}_{P_+}[c(X)w(f(X))] - \pi_+ \mathbb{E}_{P_+}[c(X)w(f(X))f(X)] \\ &\quad - \pi_- \mathbb{E}_{P_-}[c(X)w(f(X))f(X)] \\ &= \pi_+ \mathbb{E}_{P_+}[c(X)w(f(X))] - \mathbb{E}_{P_X}[c(X)w(f(X))f(X)]. \end{aligned}$$

Subtracting this identity from the definition of $R^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+)$ gives the bias identity. Finally,

$$\left|\mathbb{E}_{P_+}[c(X)w(f(X))]\right| \le \mathbb{E}_{P_+}[|c(X)w(f(X))|] \le 1$$

implies Eq. (28). □

Corollary 4 (PU nuisance and sampling errors)**.** *Assume that $\mathcal{C}$ and $\mathcal{W}$ are nonempty and finite, and let $\{X_i^+\}_{i=1}^{n_+} \overset{\text{i.i.d.}}{\sim} P_+$ and $\{X_j^u\}_{j=1}^{n_u} \overset{\text{i.i.d.}}{\sim} P_X$ be independent samples. Assume that $f, \mathcal{C}, \mathcal{W}, \widehat{\pi}_+$ are fixed independently of these samples, and define*

$$\widehat{R}^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+) := \frac{\widehat{\pi}_+}{n_+} \sum_{i=1}^{n_+} c(X_i^+)w(f(X_i^+)) - \frac{1}{n_u} \sum_{j=1}^{n_u} c(X_j^u)w(f(X_j^u))f(X_j^u).$$

For $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left| \widehat{R}^{\mathrm{PU}}_{c,w}(f;\widehat{\pi}_+) - R_{c,w}(f) \right| \le e_\pi + \widehat{\pi}_+ \sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_+}} + \sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_u}}. \tag{29}$$

Proof. Since $|c\,w(f)| \le 1$ and $|c\,w(f)f| \le 1$, the summands from both sources lie in $[-1,1]$. Applying Hoeffding's inequality to each witness and each source, and taking a union bound over $2|\mathcal{C}||\mathcal{W}|$ events, gives, with probability at least $1-\delta$,

$$\left| \frac{1}{n_+} \sum_{i=1}^{n_+} c(X_i^+)w(f(X_i^+)) - \mathbb{E}_{P_+}[c(X)w(f(X))] \right| \le \sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_+}},$$

$$\left| \frac{1}{n_u} \sum_{j=1}^{n_u} c(X_j^u)w(f(X_j^u))f(X_j^u) - \mathbb{E}_{P_X}[c(X)w(f(X))f(X)] \right| \le \sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_u}}$$

simultaneously for all $(c,w) \in \mathcal{C} \times \mathcal{W}$. Multiplying the first bound by $\widehat{\pi}_+$, adding the second bound, and then adding the bias e_π from Proposition 4 proves Eq. (29). □

D.2 UU: estimated class prior and mixture proportions

The true observation distributions are

$$P_{U_1} = (1-\gamma_1)P_+ + \gamma_1 P_-, \qquad P_{U_2} = \gamma_2 P_+ + (1-\gamma_2)P_-,$$

where $\gamma_1, \gamma_2 \in [0,1]$ and $\Delta := 1-\gamma_1-\gamma_2 \neq 0$. Given an estimate $\widehat{\pi}_+ \in [0,1]$ of the class prior π_+ and estimates $\widehat{\gamma}_1, \widehat{\gamma}_2 \in [0,1]$ of the mixture proportions γ_1, γ_2, let

$$\widehat{\gamma} := (\widehat{\gamma}_1, \widehat{\gamma}_2), \qquad \widehat{\Delta} := 1-\widehat{\gamma}_1-\widehat{\gamma}_2 \neq 0,$$

and define the estimation errors by

$$\delta_\pi := \widehat{\pi}_+ - \pi_+, \qquad \delta_k := \widehat{\gamma}_k - \gamma_k \quad (k=1,2), \qquad e_\gamma := |\delta_1| + |\delta_2|.$$

For $v \in [0,1]$, the correction coefficients using these estimates are

$$\widehat{\alpha}_1(v) := \frac{(1-\widehat{\gamma}_2)\widehat{\pi}_+(1-v) + \widehat{\gamma}_2\widehat{\pi}_- v}{\widehat{\Delta}}, \qquad \widehat{\alpha}_2(v) := -\frac{\widehat{\gamma}_1\widehat{\pi}_+(1-v) + (1-\widehat{\gamma}_1)\widehat{\pi}_- v}{\widehat{\Delta}}. \tag{30}$$

Under the true observation distributions, the population residual using these coefficients is

$$R^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) := \mathbb{E}_{P_{U_1}}[c(X)w(f(X))\widehat{\alpha}_1(f(X))] + \mathbb{E}_{P_{U_2}}[c(X)w(f(X))\widehat{\alpha}_2(f(X))]. \tag{31}$$

Proposition 5 (UU nuisance error)**.** *Under the assumptions above,*

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left| R^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) - R_{c,w}(f) \right| \leq 2e_\pi + \frac{2e_\gamma}{|\widehat{\Delta}|}. \tag{32}$$

If $e_\gamma \leq |\Delta|/2$, the right-hand side is at most $2e_\pi + 4e_\gamma/|\Delta|$. If $|\Delta|, |\widehat{\Delta}| \geq \Delta_{\min} > 0$, it is at most $2e_\pi + 2e_\gamma/\Delta_{\min}$.

Proof. The coefficients in Eq. (30) satisfy

$$\begin{aligned} (1-\widehat{\gamma}_1)\widehat{\alpha}_1(v) + \widehat{\gamma}_2\widehat{\alpha}_2(v) &= \widehat{\pi}_+(1-v), \\ \widehat{\gamma}_1\widehat{\alpha}_1(v) + (1-\widehat{\gamma}_2)\widehat{\alpha}_2(v) &= -\widehat{\pi}_- v. \end{aligned} \tag{33}$$

Indeed, the inverse of the coefficient matrix is

$$\begin{pmatrix} 1-\widehat{\gamma}_1 & \widehat{\gamma}_2 \\ \widehat{\gamma}_1 & 1-\widehat{\gamma}_2 \end{pmatrix}^{-1} = \frac{1}{\widehat{\Delta}} \begin{pmatrix} 1-\widehat{\gamma}_2 & -\widehat{\gamma}_2 \\ -\widehat{\gamma}_1 & 1-\widehat{\gamma}_1 \end{pmatrix}.$$

Multiplying this matrix by $(\widehat{\pi}_+(1-v), -\widehat{\pi}_- v)^\top$ gives Eq. (30).

Set $d(v) := \delta_1\widehat{\alpha}_1(v) - \delta_2\widehat{\alpha}_2(v)$. Substituting $\gamma_k = \widehat{\gamma}_k - \delta_k$ into Eq. (33) gives

$$\begin{aligned} (1-\gamma_1)\widehat{\alpha}_1(v) + \gamma_2\widehat{\alpha}_2(v) &= (1-\widehat{\gamma}_1+\delta_1)\widehat{\alpha}_1(v) + (\widehat{\gamma}_2-\delta_2)\widehat{\alpha}_2(v) \\ &= \widehat{\pi}_+(1-v) + d(v) = \pi_+(1-v) + \delta_\pi(1-v) + d(v), \\ \gamma_1\widehat{\alpha}_1(v) + (1-\gamma_2)\widehat{\alpha}_2(v) &= (\widehat{\gamma}_1-\delta_1)\widehat{\alpha}_1(v) + (1-\widehat{\gamma}_2+\delta_2)\widehat{\alpha}_2(v) \\ &= -\widehat{\pi}_- v - d(v) = -\pi_- v + \delta_\pi v - d(v), \end{aligned}$$

where the last equality uses $\widehat{\pi}_- = \pi_- - \delta_\pi$. Expanding the expectations in Eq. (31) under the class-conditional distributions yields

$$\begin{aligned} R^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) &= (1-\gamma_1)\mathbb{E}_{P_+}[c(X)w(f(X))\widehat{\alpha}_1(f(X))] + \gamma_1\mathbb{E}_{P_-}[c(X)w(f(X))\widehat{\alpha}_1(f(X))] \\ &\quad + \gamma_2\mathbb{E}_{P_+}[c(X)w(f(X))\widehat{\alpha}_2(f(X))] + (1-\gamma_2)\mathbb{E}_{P_-}[c(X)w(f(X))\widehat{\alpha}_2(f(X))] \\ &= \mathbb{E}_{P_+}[c(X)w(f(X))\{\pi_+(1-f(X)) + \delta_\pi(1-f(X)) + d(f(X))\}] \\ &\quad + \mathbb{E}_{P_-}[c(X)w(f(X))\{-\pi_- f(X) + \delta_\pi f(X) - d(f(X))\}]. \end{aligned}$$

Subtracting the true residual $R_{c,w}(f)$ cancels the two corresponding terms, giving

$$\begin{aligned} R^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) - R_{c,w}(f) &= \mathbb{E}_{P_+}[c(X)w(f(X))\{\delta_\pi(1-f(X)) + d(f(X))\}] \\ &\quad + \mathbb{E}_{P_-}[c(X)w(f(X))\{\delta_\pi f(X) - d(f(X))\}]. \end{aligned} \tag{34}$$

Apart from the leading minus sign, each numerator in Eq. (30) is a convex combination of two values in $[0,1]$. For example, the first numerator combines $(1-\widehat{\gamma}_2)\widehat{\pi}_+$ and $\widehat{\gamma}_2\widehat{\pi}_-$ with weights $1-v$ and v, respectively. Therefore,

$$|\widehat{\alpha}_k(v)| \le \frac{1}{|\widehat{\Delta}|} \quad (k=1,2), \qquad |d(v)| \le |\delta_1|\,|\widehat{\alpha}_1(v)| + |\delta_2|\,|\widehat{\alpha}_2(v)| \le \frac{e_\gamma}{|\widehat{\Delta}|}.$$

Using these bounds, $|c\,w(f)| \le 1$, and $0 \le f \le 1$ in Eq. (34) gives

$$\begin{aligned}\left|R^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) - R_{c,w}(f)\right| &\le e_\pi\{\mathbb{E}_{P_+}[1-f(X)] + \mathbb{E}_{P_-}[f(X)]\} \\ &\quad + \mathbb{E}_{P_+}[|d(f(X))|] + \mathbb{E}_{P_-}[|d(f(X))|] \le 2e_\pi + \frac{2e_\gamma}{|\widehat{\Delta}|}.\end{aligned}$$

The bound is independent of (c,w), proving Eq. (32). Finally,

$$|\widehat{\Delta} - \Delta| = |\delta_1 + \delta_2| \le e_\gamma, \qquad |\widehat{\Delta}| \ge |\Delta| - e_\gamma.$$

Thus, $e_\gamma \le |\Delta|/2$ implies $|\widehat{\Delta}| \ge |\Delta|/2$. Substituting the lower bounds on the denominator into Eq. (32) proves the remaining two bounds. □

Corollary 5 (UU nuisance and sampling errors)**.** *Assume that $\mathcal{C}$ and $\mathcal{W}$ are nonempty and finite, and for $k=1,2$, let $\{X_i^{U_k}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} P_{U_k}$. Assume that $f, \mathcal{C}, \mathcal{W}, \widehat{\pi}_+, \widehat{\gamma}_1, \widehat{\gamma}_2$ are fixed independently of these samples, and define*

$$\widehat{R}^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) := \sum_{k=1}^{2} \frac{1}{n_k} \sum_{i=1}^{n_k} c(X_i^{U_k}) w(f(X_i^{U_k})) \widehat{\alpha}_k(f(X_i^{U_k})).$$

For $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left|\widehat{R}^{\mathrm{UU}}_{c,w}(f;\widehat{\pi}_+,\widehat{\gamma}) - R_{c,w}(f)\right| \le 2e_\pi + \frac{2e_\gamma}{|\widehat{\Delta}|} + \frac{1}{|\widehat{\Delta}|}\sum_{k=1}^{2}\sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_k}}. \tag{35}$$

Proof. Set $Z_{k,c,w}(x) := c(x)w(f(x))\widehat{\alpha}_k(f(x))$. Then $|Z_{k,c,w}| \le 1/|\widehat{\Delta}|$. Hoeffding's inequality gives

$$\Pr\left\{\left|\frac{1}{n_k}\sum_{i=1}^{n_k} Z_{k,c,w}(X_i^{U_k}) - \mathbb{E}_{P_{U_k}}[Z_{k,c,w}(X)]\right| \ge t\right\} \le 2\exp\left(-\frac{n_k|\widehat{\Delta}|^2 t^2}{2}\right).$$

Choose

$$t = \frac{1}{|\widehat{\Delta}|}\sqrt{\frac{2\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{n_k}}$$

and take a union bound over the two distributions and $|\mathcal{C}||\mathcal{W}|$ witnesses. Adding the resulting errors from the two sources and the bias in Proposition 5 proves Eq. (35). □

With the true parameters and $n_1 = n_2 = n$,

$$n \ge \frac{8\log(4|\mathcal{C}||\mathcal{W}|/\delta)}{\Delta^2\varepsilon^2}$$

is sufficient to make the sampling error at most ε. Moreover, $P_{U_1} - P_{U_2} = \Delta(P_+ - P_-)$, so when $\Delta = 0$, the two observation distributions coincide and this system cannot separate P_+ and P_-.

D.3 Pconf: estimated confidences and clipping

Let $r(x) := P(Y=1 \mid X=x)$, and assume that $r(X) > 0$ holds P_X-almost surely. By Bayes' rule, $\pi_+\,dP_+ = r\,dP_X$. Hence,

$$R_{c,w}(f) = \mathbb{E}_{P_X}[c(X)w(f(X))\{r(X) - f(X)\}] = \pi_+\mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{r(X)}\right)\right]. \tag{36}$$

The first equality follows by conditioning on X, and the second follows from the measure identity above. Moreover,

$$\pi_+\mathbb{E}_{P_+}\left[|c(X)w(f(X))|\left|1 - \frac{f(X)}{r(X)}\right|\right] = \mathbb{E}_{P_X}[|c(X)w(f(X))|\,|r(X) - f(X)|] \le 1,$$

so the integrand in the last expectation in Eq. (36) is integrable.

Parameter estimation. Given an estimate $\widehat{\pi}_+ \in [0,1]$ of the class prior π_+ and an estimate $\widehat{r} : \mathcal{X} \to [0,1]$ of the confidence function r, let $e_r := \|\widehat{r} - r\|_{L^\infty(P_X)}$. Assume that, for some $C_r \in (0,1]$,

$$r(X) \geq C_r, \qquad \widehat{r}(X) \geq C_r \qquad P_X\text{-almost surely.}$$

No clipping is applied here. Define the population residual using the estimates by

$$R^{\mathrm{Pconf}}_{c,w}(f; \widehat{\pi}_+, \widehat{r}) := \widehat{\pi}_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{\widehat{r}(X)}\right)\right].$$

Proposition 6. *Under the assumptions above,*

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} \left|R^{\mathrm{Pconf}}_{c,w}(f; \widehat{\pi}_+, \widehat{r}) - R_{c,w}(f)\right| \leq \frac{e_\pi}{C_r} + \frac{\pi_+ e_r}{C_r^2}.$$

Proof. Add and subtract the population residual using π_+ and $\widehat{r}$. By Eq. (36),

$$\begin{aligned} R^{\mathrm{Pconf}}_{c,w}(f; \widehat{\pi}_+, \widehat{r}) - R_{c,w}(f) = (\widehat{\pi}_+ - \pi_+)\mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{\widehat{r}(X)}\right)\right] \\ + \pi_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))f(X)\left(\frac{1}{r(X)} - \frac{1}{\widehat{r}(X)}\right)\right]. \end{aligned}$$

For the first term, $0 \leq f \leq 1$ and $\widehat{r} \geq C_r$ imply

$$0 \leq \frac{f(X)}{\widehat{r}(X)} \leq C_r^{-1}, \qquad 1 - C_r^{-1} \leq 1 - \frac{f(X)}{\widehat{r}(X)} \leq 1.$$

Comparing the absolute values of the two endpoints of this interval gives

$$\left|1 - \frac{f(X)}{\widehat{r}(X)}\right| \leq \max\{1, C_r^{-1} - 1\} \leq C_r^{-1}.$$

The last inequality follows because $0 < C_r \leq 1$ implies $1 \leq C_r^{-1}$ and $C_r^{-1} - 1 \leq C_r^{-1}$. For the second term,

$$\left|\frac{1}{r(X)} - \frac{1}{\widehat{r}(X)}\right| = \frac{|\widehat{r}(X) - r(X)|}{r(X)\widehat{r}(X)} \leq \frac{e_r}{C_r^2}.$$

Therefore, using $|c\, w(f)| \leq 1$ and $0 \leq f \leq 1$, we obtain

$$\begin{aligned} \left|R^{\mathrm{Pconf}}_{c,w}(f; \widehat{\pi}_+, \widehat{r}) - R_{c,w}(f)\right| &\leq e_\pi \mathbb{E}_{P_+}\left[\left|1 - \frac{f(X)}{\widehat{r}(X)}\right|\right] + \pi_+ \mathbb{E}_{P_+}\left[f(X)\left|\frac{1}{r(X)} - \frac{1}{\widehat{r}(X)}\right|\right] \\ &\leq \frac{e_\pi}{C_r} + \frac{\pi_+ e_r}{C_r^2}. \end{aligned}$$

Taking the supremum over (c, w) proves the claim. □

Clipping bias. Here we use the true values of π_+ and r. Fix $\tau \in (0,1]$ and define

$$R^{\mathrm{Pconf},\tau}_{c,w}(f) := \pi_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{\tau \vee r(X)}\right)\right].$$

In this part, we do not assume a positive uniform lower bound on r. We continue to assume that $r > 0$ holds P_X-almost surely.

Proposition 7 (Clipping bias)**.** *The signed bias is*

$$R^{\mathrm{Pconf},\tau}_{c,w}(f) - R_{c,w}(f) = \mathbb{E}_{P_X}\left[c(X)w(f(X))f(X)\left(1 - \frac{r(X)}{\tau}\right)\mathbf{1}\{r(X) < \tau\}\right]. \tag{37}$$

In particular,

$$\begin{aligned} \sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |R^{\mathrm{Pconf},\tau}_{c,w}(f) - R_{c,w}(f)| \leq B_\tau(f) &:= \mathbb{E}_{P_X}\left[f(X)\left(1 - \frac{r(X)}{\tau}\right)\mathbf{1}\{r(X) < \tau\}\right] \\ &\leq P_X\{r(X) < \tau\}. \end{aligned}$$

Moreover, $B_\tau(f)$ is nondecreasing in τ, and $B_\tau(f) \to 0$ as $\tau \downarrow 0$.

Proof. First, subtracting Eq. (36) from $R_{c,w}^{\mathrm{Pconf},\tau}(f)$ gives

$$R_{c,w}^{\mathrm{Pconf},\tau}(f) - R_{c,w}(f) = \pi_+ \mathbb{E}_{P_+}\left[c(X)w(f(X))f(X)\left(\frac{1}{r(X)} - \frac{1}{\tau \vee r(X)}\right)\right].$$

On $\{r \geq \tau\}$, we have $\tau \vee r = r$, so the difference is zero; on $\{r < \tau\}$, we have $\tau \vee r = \tau$. Thus, only the integral over $\{x : r(x) < \tau\}$ remains. Substituting Bayes' rule $\pi_+ P_+(dx) = r(x)P_X(dx)$ yields

$$\begin{aligned} R_{c,w}^{\mathrm{Pconf},\tau}(f) - R_{c,w}(f) &= \pi_+ \int_{\{r<\tau\}} c(x)w(f(x))f(x)\left(\frac{1}{r(x)} - \frac{1}{\tau}\right) P_+(dx) \\ &= \int_{\{r<\tau\}} c(x)w(f(x))f(x)r(x)\left(\frac{1}{r(x)} - \frac{1}{\tau}\right) P_X(dx) \\ &= \mathbb{E}_{P_X}\left[c(X)w(f(X))f(X)\left(1 - \frac{r(X)}{\tau}\right)\mathbf{1}\{r(X) < \tau\}\right]. \end{aligned}$$

This proves Eq. (37). Since $1 - r/\tau \geq 0$ on $\{r < \tau\}$, taking absolute values and using $|c\, w(f)| \leq 1$ gives

$$|R_{c,w}^{\mathrm{Pconf},\tau}(f) - R_{c,w}(f)| \leq \mathbb{E}_{P_X}\left[f(X)\left(1 - \frac{r(X)}{\tau}\right)\mathbf{1}\{r(X) < \tau\}\right] = B_\tau(f).$$

Furthermore, $0 \leq f(1 - r/\tau)\mathbf{1}\{r < \tau\} \leq \mathbf{1}\{r < \tau\}$ implies $B_\tau(f) \leq P_X\{r(X) < \tau\}$.

This bias bound has the following integral representation over regions with small confidence values. For fixed $u > 0$, the integral $\int_0^\tau \mathbf{1}\{u \leq s\}\, ds$ is zero if $u \geq \tau$, and equals $\tau - u$ if $u < \tau$. Therefore,

$$\left(1 - \frac{u}{\tau}\right)\mathbf{1}\{u < \tau\} = \frac{1}{\tau}\int_0^\tau \mathbf{1}\{u \leq s\}\, ds.$$

Substituting $u = r(X)$ and interchanging the expectation and integral, which is valid because the integrand is nonnegative, gives

$$\begin{aligned} B_\tau(f) &= \mathbb{E}_{P_X}\left[f(X)\frac{1}{\tau}\int_0^\tau \mathbf{1}\{r(X) \leq s\}\, ds\right] \\ &= \frac{1}{\tau}\int_0^\tau \mathbb{E}_{P_X}[f(X)\mathbf{1}\{r(X) \leq s\}]\, ds \leq \frac{1}{\tau}\int_0^\tau P_X\{r(X) \leq s\}\, ds \leq P_X\{r(X) < \tau\}. \end{aligned}$$

The last inequality uses $\{r \leq s\} \subset \{r < \tau\}$ for $s < \tau$.

For every $u > 0$, the function $(1 - u/\tau)\mathbf{1}\{u < \tau\}$ is zero when $\tau \leq u$ and increases with τ when $\tau > u$. Multiplying by $f \geq 0$ and taking the expectation proves that $B_\tau(f)$ is nondecreasing. Finally, the integrand defining $B_\tau(f)$ converges to zero as $\tau \downarrow 0$ and is bounded by one. The claimed limit therefore follows from the dominated convergence theorem. □

Corollary 6 (Clipping bias and sampling error)**.** *Assume that $\mathcal{C}$ and $\mathcal{W}$ are finite. Let $f, \mathcal{C}, \mathcal{W}, \tau$ be fixed independently of the sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_+$. Define the empirical residual by*

$$\widehat{R}_{c,w}^{\mathrm{Pconf},\tau}(f) := \frac{\pi_+}{n}\sum_{i=1}^n c(X_i)w(f(X_i))\left(1 - \frac{f(X_i)}{\tau \vee r(X_i)}\right).$$

For $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}} |\widehat{R}_{c,w}^{\mathrm{Pconf},\tau}(f) - R_{c,w}(f)| \leq B_\tau(f) + \frac{\pi_+}{\tau}\sqrt{\frac{2\log(2|\mathcal{C}||\mathcal{W}|/\delta)}{n}}. \tag{38}$$

Proof. Let

$$Z_{\tau,c,w}(X) := \pi_+ c(X)w(f(X))\left(1 - \frac{f(X)}{\tau \vee r(X)}\right).$$

Since $0 \leq f/(\tau \vee r) \leq 1/\tau$ and $0 < \tau \leq 1$,

$$|Z_{\tau,c,w}(X)| \leq \pi_+ \max\{1, \tau^{-1} - 1\} \leq \frac{\pi_+}{\tau}.$$

The expectation of this random variable is $R^{\mathrm{Pconf},\tau}_{c,w}(f)$, and its range is contained in an interval of length at most $2\pi_+/\tau$. Thus, Hoeffding's inequality gives

$$\Pr\left\{|\widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f) - R^{\mathrm{Pconf},\tau}_{c,w}(f)| \geq t\right\} \leq 2\exp\left(-\frac{n\tau^2 t^2}{2\pi_+^2}\right).$$

Set $t = (\pi_+/\tau)\sqrt{2\log(2|\mathcal{C}||\mathcal{W}|/\delta)/n}$ and sum the failure probabilities over the $|\mathcal{C}||\mathcal{W}|$ pairs (c, w). On the resulting event of probability at least $1-\delta$, the triangle inequality and Proposition 7 imply

$$\begin{aligned}\sup_{c,w}|\widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f) - R_{c,w}(f)| &\leq \sup_{c,w}|R^{\mathrm{Pconf},\tau}_{c,w}(f) - R_{c,w}(f)| + \sup_{c,w}|\widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f) - R^{\mathrm{Pconf},\tau}_{c,w}(f)| \\ &\leq B_\tau(f) + \frac{\pi_+}{\tau}\sqrt{\frac{2\log(2|\mathcal{C}||\mathcal{W}|/\delta)}{n}}.\end{aligned}$$

□

In Eq. (38), the bias bound is nondecreasing in τ, whereas the Hoeffding term decreases in proportion to $1/\tau$. This is the tradeoff between clipping bias and sampling error.

Clipping estimated confidences. Given an estimate $\widehat{\pi}_+ \in [0,1]$ of the class prior π_+ and an estimate $\widehat{r} : \mathcal{X} \to [0,1]$ of the confidence function r, let $e_r = \|\widehat{r} - r\|_{L^\infty(P_X)}$ and clip it as $\tau \vee \widehat{r}$. Here we do not assume positive uniform lower bounds on r or $\widehat{r}$. Define the population and empirical residuals, respectively, by

$$\begin{aligned}R^{\mathrm{Pconf},\tau}_{c,w}(f;\widehat{\pi}_+,\widehat{r}) &:= \widehat{\pi}_+\mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{\tau \vee \widehat{r}(X)}\right)\right], \\ \widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f;\widehat{\pi}_+,\widehat{r}) &:= \frac{\widehat{\pi}_+}{n}\sum_{i=1}^{n} c(X_i)w(f(X_i))\left(1 - \frac{f(X_i)}{\tau \vee \widehat{r}(X_i)}\right).\end{aligned}$$

Corollary 7. *The population bias is bounded by*

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}}|R^{\mathrm{Pconf},\tau}_{c,w}(f;\widehat{\pi}_+,\widehat{r}) - R_{c,w}(f)| \leq B_\tau(f) + \frac{e_\pi}{\tau} + \frac{\pi_+ e_r}{\tau^2}. \tag{39}$$

Furthermore, suppose that $\mathcal{C}$ and $\mathcal{W}$ are finite and $f, \mathcal{C}, \mathcal{W}, \tau, \widehat{\pi}_+, \widehat{r}$ are fixed independently of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_+$. For $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{c\in\mathcal{C},\, w\in\mathcal{W}}|\widehat{R}^{\mathrm{Pconf},\tau}_{c,w}(f;\widehat{\pi}_+,\widehat{r}) - R_{c,w}(f)| \leq B_\tau(f) + \frac{e_\pi}{\tau} + \frac{\pi_+ e_r}{\tau^2} + \frac{\widehat{\pi}_+}{\tau}\sqrt{\frac{2\log(2|\mathcal{C}||\mathcal{W}|/\delta)}{n}}. \tag{40}$$

Proof. Since $|\max\{u,\tau\} - \max\{v,\tau\}| \leq |u-v|$,

$$\left|\frac{1}{\tau \vee \widehat{r}} - \frac{1}{\tau \vee r}\right| = \frac{|(\tau \vee \widehat{r}) - (\tau \vee r)|}{(\tau \vee \widehat{r})(\tau \vee r)} \leq \frac{|\widehat{r} - r|}{\tau^2} \leq \frac{e_r}{\tau^2} \qquad P_X\text{-almost surely.}$$

Comparing with the residual $R^{\mathrm{Pconf},\tau}_{c,w}(f)$ obtained by clipping the true confidence gives

$$\begin{aligned}R^{\mathrm{Pconf},\tau}_{c,w}(f;\widehat{\pi}_+,\widehat{r}) - R^{\mathrm{Pconf},\tau}_{c,w}(f) &= (\widehat{\pi}_+ - \pi_+)\mathbb{E}_{P_+}\left[c(X)w(f(X))\left(1 - \frac{f(X)}{\tau \vee \widehat{r}(X)}\right)\right] \\ &\quad + \pi_+\mathbb{E}_{P_+}\left[c(X)w(f(X))f(X)\left(\frac{1}{\tau \vee r(X)} - \frac{1}{\tau \vee \widehat{r}(X)}\right)\right].\end{aligned}$$

The absolute value of the first term is at most e_π/τ, and that of the second is at most $\pi_+ e_r/\tau^2$. Adding the bias in Proposition 7 gives Eq. (39). Each summand in the empirical average has absolute value at most $\widehat{\pi}_+/\tau$, so adding the same Hoeffding bound as in Corollary 6 yields Eq. (40). □

E Proofs for Section 5

E.1 Fresh-sample version of WLMC

Algorithm 2 draws new independent weak samples at each search call, whereas Algorithm 1 in the main paper draws the samples once and reuses them throughout.

Algorithm 2 Fresh-sample weak-label multicalibration boost (WLMC)

Input: $\varepsilon > 0$, source-wise sample sizes $\mathbf{n}$, step size λ, initial predictor f_0, classes $\mathcal{C}, \mathcal{W}$, estimator $\widehat{R}^{\mathrm{WSL}}$
Set $t = 0$ and draw a fresh weak batch S_0 with source-wise sizes $\mathbf{n}$
while $\exists(\tau, c, w) \in \{\pm 1\} \times \mathcal{C} \times \mathcal{W}$ such that $\tau \widehat{R}^{\mathrm{WSL}}_{c,w}(f_t; S_t) > 3\varepsilon/4$ **do**
  Choose $(\tau_t, c_t, w_t) \in \{\pm 1\} \times \mathcal{C} \times \mathcal{W}$ satisfying the while condition
  $\forall x \in \mathcal{X}, \quad f_{t+1}(x) = \Pi_{[0,1]}[f_t(x) + \lambda \tau_t c_t(x) w_t(f_t(x))]$
  $t \leftarrow t + 1$
  Draw a fresh weak batch S_t with source-wise sizes $\mathbf{n}$, independently of the past
end while
Output: f_t

Theorem 3 (Formal PU guarantee under Rademacher complexity)**.** *Let $0 < \varepsilon \leq 1$ and $0 < \delta < 1$. Assume the PU setting of Proposition 1 with known class prior π_+, and fix $f_0, \mathcal{C}, \mathcal{W}$ independently of the calibration samples. Suppose $|c| \leq 1$, $\|w\|_\infty \leq 1$, and $\mathcal{W}$ is finite. Assume further that the Rademacher complexities are bounded as*

$$\mathfrak{R}_{n_+}(\mathcal{C}; P_+) \leq C_R \sqrt{\frac{d_{\mathcal{C}} \log n_+}{n_+}}, \qquad \mathfrak{R}_{n_u}(\mathcal{C}; P_X) \leq C_R \sqrt{\frac{d_{\mathcal{C}} \log n_u}{n_u}}.$$

Let $D_0 = \mathbb{E}[(f_0(X) - f^\star(X))^2]$ and $T_\varepsilon = \lceil 4D_0/\varepsilon^2 \rceil$. Run Algorithm 2 with $\lambda = \varepsilon/2$, threshold $3\varepsilon/4$, and the PU estimator $\widehat{R}^{\mathrm{PU}}$. At each search call, draw fresh independent batches $S_t^+ \sim P_+^{n_+}$ and $S_t^u \sim P_X^{n_u}$. Let

$$L = \log \frac{4|\mathcal{W}|(T_\varepsilon + 1)}{\delta}.$$

There exists $C > 0$, depending only on C_R, such that if

$$n_+ \geq \frac{C\pi_+^2}{\varepsilon^2}(d_{\mathcal{C}} \log n_+ + L), \qquad n_u \geq \frac{C}{\varepsilon^2}(d_{\mathcal{C}} \log n_u + L),$$

then, with probability at least $1 - \delta$, Algorithm 2 stops after $\widehat{T} \leq T_\varepsilon$ successful updates and returns a $(\mathcal{C}, \mathcal{W}, \varepsilon)$-multicalibrated predictor $f_{\widehat{T}}$. In particular, if $n_+ = n_u = n$, it suffices that

$$n \geq \frac{C}{\varepsilon^2}\left[d_{\mathcal{C}} \log n + \log|\mathcal{W}| + \log \frac{4(T_\varepsilon + 1)}{\delta}\right].$$

On the same event, the total source-wise budgets are $N_+ \leq (T_\varepsilon + 1)n_+$ and $N_u \leq (T_\varepsilon + 1)n_u$.

Proof. The proof is similar to Gopalan et al. [17]. However, we need to control the finite-sample estimation error.

For a fixed predictor f, define the population residual moment

$$R^{\mathrm{WSL}}_{c,w}(f) = \mathbb{E}[c(X)w(f(X))(Y - f(X))] = \mathbb{E}[c(X)w(f(X))(f^\star(X) - f(X))].$$

In the PU setting, the same moment is represented as

$$R^{\mathrm{PU}}_{c,w}(f) = \pi_+ \mathbb{E}_{P_+}[c(X)w(f(X))] - \mathbb{E}_{P_X}[c(X)w(f(X))f(X)].$$

Consider a population search rule which either returns $(\tau, c, w) \in \{\pm 1\} \times \mathcal{C} \times \mathcal{W}$ with $\tau R^{\mathrm{WSL}}_{c,w}(f) \geq \varepsilon/2$, or returns $\perp$, in which case $\sup_{c,w} |R^{\mathrm{WSL}}_{c,w}(f)| \leq \varepsilon$.

Let

$$\Psi(f) = \mathbb{E}[(f(X) - f^\star(X))^2].$$

If the population rule returns (τ_t, c_t, w_t), the update is

$$f_{t+1}(x) = \Pi_{[0,1]}\left[f_t(x) + \frac{\varepsilon}{2}\tau_t c_t(x) w_t(f_t(x))\right].$$

Set

$$\Delta_t(x) := \tau_t c_t(x) w_t(f_t(x)).$$

Since $f^\star(x) \in [0,1]$ and projection onto $[0,1]$ is non-expansive,

$$(f_{t+1}(x) - f^\star(x))^2 \le \left(f_t(x) + \frac{\varepsilon}{2}\Delta_t(x) - f^\star(x)\right)^2.$$

Taking expectations and expanding the square, we obtain

$$\begin{aligned}
\Psi(f_{t+1}) &\le \mathbb{E}\left[\left(f_t(X) + \frac{\varepsilon}{2}\Delta_t(X) - f^\star(X)\right)^2\right] \\
&= \Psi(f_t) + \varepsilon\mathbb{E}[\Delta_t(X)(f_t(X) - f^\star(X))] + \frac{\varepsilon^2}{4}\mathbb{E}[\Delta_t(X)^2] \\
&= \Psi(f_t) - \varepsilon\mathbb{E}[\Delta_t(X)(f^\star(X) - f_t(X))] + \frac{\varepsilon^2}{4}\mathbb{E}[\Delta_t(X)^2] \\
&= \Psi(f_t) - \varepsilon\tau_t R^{\mathrm{WSL}}_{c_t,w_t}(f_t) + \frac{\varepsilon^2}{4}\mathbb{E}[\Delta_t(X)^2].
\end{aligned}$$

Since $|c_t| \le 1$, $\|w_t\|_\infty \le 1$, and $\tau_t \in \{\pm 1\}$, we have $|\Delta_t| \le 1$. Moreover, the population witness satisfies

$$\tau_t R^{\mathrm{WSL}}_{c_t,w_t}(f_t) \ge \varepsilon/2.$$

Therefore,

$$\Psi(f_{t+1}) \le \Psi(f_t) - \varepsilon \cdot \frac{\varepsilon}{2} + \frac{\varepsilon^2}{4} = \Psi(f_t) - \frac{\varepsilon^2}{4}.$$

Thus every successful population update decreases the potential by at least $\varepsilon^2/4$. After M successful updates,

$$\Psi(f_M) \le D_0 - \frac{M\varepsilon^2}{4}.$$

Since $\Psi(f) \ge 0$ for every predictor f, we must have

$$M \le \frac{4D_0}{\varepsilon^2}.$$

Otherwise the potential would become negative, which is impossible. Hence the population procedure cannot keep finding valid update directions forever; after at most $4D_0/\varepsilon^2$ successful updates, it must fail to find another valid witness and therefore stops. In particular, the number of successful population updates is at most $T_\varepsilon = \lceil 4D_0/\varepsilon^2 \rceil$.

Next, we control the finite-sample estimation error. At each reached search call t, condition on the history before drawing the new batches. Then f_t is fixed, and S_t^+ and S_t^u are independent of that history. For $t = 0, \ldots, T_\varepsilon$, let $\mathcal{E}_t$ be the event that either call t is not reached or its residual estimation error satisfies

$$\sup_{c\in\mathcal{C}, w\in\mathcal{W}} \left|\widehat{R}^{\mathrm{PU}}_{c,w}(f_t; S_t^+, S_t^u) - R^{\mathrm{PU}}_{c,w}(f_t)\right| \le \varepsilon/4.$$

On $\mathcal{E}_t$, if the empirical search returns a witness with

$$\tau_t \widehat{R}^{\mathrm{PU}}_{c_t,w_t}(f_t) > 3\varepsilon/4,$$

then

$$\tau_t R^{\mathrm{PU}}_{c_t,w_t}(f_t) > \varepsilon/2.$$

Hence the update is a valid population update and decreases Ψ by at least $\varepsilon^2/4$. Conversely, if the empirical search stops, then

$$\sup_{c,w} |\widehat{R}^{\mathrm{PU}}_{c,w}(f_t)| \le 3\varepsilon/4,$$

and therefore $\sup_{c,w} |R^{\mathrm{PU}}_{c,w}(f_t)| \le \varepsilon$. Thus the output is $(\mathcal{C}, \mathcal{W}, \varepsilon)$-multicalibrated.

Apply the finite-$\mathcal{W}$ residual-process bound established in the proof of Proposition 1 with failure probability $\delta/(T_\varepsilon+1)$. Using the assumed Rademacher bound, there exists $C>0$, depending only on C_R, such that the conditional failure probability at every reached call is at most $\delta/(T_\varepsilon+1)$. Since $\mathcal{E}_t$ holds automatically after stopping, averaging over the past gives $\Pr(\mathcal{E}_t^c) \le \delta/(T_\varepsilon+1)$ whenever

$$n_+ \ge \frac{C\pi_+^2}{\varepsilon^2}(d_{\mathcal{C}} \log n_+ + L), \qquad n_u \ge \frac{C}{\varepsilon^2}(d_{\mathcal{C}} \log n_u + L),$$

where $L = \log(4|\mathcal{W}|(T_\varepsilon+1)/\delta)$.

By a union bound over the first $T_\varepsilon+1$ search calls,

$$\Pr\left(\bigcap_{t=0}^{T_\varepsilon} \mathcal{E}_t\right) \ge 1-\delta.$$

On this event, every empirical update at a reached call $t \le T_\varepsilon$ is a valid population update. Since more than T_ε successful updates are impossible, Algorithm 2 stops within the first $T_\varepsilon+1$ search calls. The sample budgets follow by multiplying the per-call source-wise sample sizes by the number of search calls:

$$N_+ \le (T_\varepsilon+1)n_+, \qquad N_u \le (T_\varepsilon+1)n_u.$$

This proves the theorem. □

E.2 Sample-reuse version of WLMC

We prove the guarantee for Algorithm 1, which draws the calibration samples once at the beginning and reuses them at every iteration. We assume that both $\mathcal{C}$ and $\mathcal{W}$ are finite, a stronger restriction than in the guarantee for Algorithm 2, which allows an infinite $\mathcal{C}$ with controlled Rademacher complexity. The proof controls the residual estimation error uniformly over all predictors reachable within a deterministic number of updates.

Theorem 4 (Formal PU guarantee for finite classes)**.** *Let $0<\varepsilon\le 1$ and $0<\delta<1$. Assume the PU setting of Proposition 1 with known class prior π_+. Suppose $\mathcal{C}$ and $\mathcal{W}$ are finite, and $|c|\le 1$, $\|w\|_\infty \le 1$. Assume that $f_0, \mathcal{C}, \mathcal{W}$ are fixed independently of the calibration samples. Let $D_0 = \mathbb{E}[(f_0(X)-f^\star(X))^2]$ and $T_\varepsilon = \lceil 4D_0/\varepsilon^2 \rceil$. Run Algorithm 1 with $\lambda = \varepsilon/2$, threshold $3\varepsilon/4$, and the PU estimator $\widehat{R}^{\mathrm{PU}}$, using independent calibration samples $S^+ \sim P_+^{n_+}$ and $S^u \sim P_X^{n_u}$ drawn once and then reused. Let $K = 2|\mathcal{C}||\mathcal{W}|$. If*

$$n_+ \ge \frac{128\pi_+^2}{\varepsilon^2}\log\frac{2K^{T_\varepsilon+2}}{\delta}, \qquad n_u \ge \frac{128}{\varepsilon^2}\log\frac{2K^{T_\varepsilon+2}}{\delta},$$

then, with probability at least $1-\delta$, Algorithm 1 stops after $\widehat{T} \le T_\varepsilon$ successful updates and returns a $(\mathcal{C},\mathcal{W},\varepsilon)$-multicalibrated predictor $f_{\widehat{T}}$.

In particular, if $n_+ = n_u = n$, it suffices that

$$n \ge \frac{128}{\varepsilon^2}\log\frac{2K^{T_\varepsilon+2}}{\delta}.$$

The total source-wise budgets are $N_+ = n_+$ and $N_u = n_u$; in the equal-size case, a sufficient budget is

$$N_+ = N_u \;=\; \mathcal{O}\left(\frac{(T_\varepsilon+2)\log K + \log(1/\delta)}{\varepsilon^2}\right).$$

Proof. For $S=(S^+,S^u)$, write $\widehat{R}^{\mathrm{PU}}_{c,w}(f;S)$ for the estimator in Eq. (12). Define the update map and the classes of reachable predictors by

$$U_{\tau,c,w}(f)(x) := \Pi_{[0,1]}\left[f(x) + \frac{\varepsilon}{2}\tau c(x) w(f(x))\right],$$

$$\mathcal{F}^{(0)} = \{f_0\},$$
$$\mathcal{F}^{(t+1)} = \left\{U_{\tau,c,w}(f) : f \in \mathcal{F}^{(t)},\ (\tau,c,w) \in \{\pm 1\}\times\mathcal{C}\times\mathcal{W}\right\}.$$

and let $\mathcal{F}_{\le T_\varepsilon} = \bigcup_{t=0}^{T_\varepsilon} \mathcal{F}^{(t)}$. These classes contain all signed update sequences, regardless of whether the empirical search would choose them. Since $f_0, \mathcal{C}, \mathcal{W}$, and the update rule are fixed independently of S, so is $\mathcal{F}_{\le T_\varepsilon}$. Every actual iterate f_t, whenever it exists and $t \le T_\varepsilon$, belongs to this class. Moreover,

$$N_{T_\varepsilon} := |\mathcal{F}_{\le T_\varepsilon}| \le \sum_{t=0}^{T_\varepsilon} K^t \le K^{T_\varepsilon+1}.$$

For a fixed triple (f, c, w), both $c(x)w(f(x))$ and $c(x)w(f(x))f(x)$ lie in $[-1, 1]$. Define

$$\mathcal{E}_{T_\varepsilon} := \left\{ \sup_{\substack{f \in \mathcal{F}_{\le T_\varepsilon} \\ c \in \mathcal{C},\ w \in \mathcal{W}}} \left| \widehat{R}^{\mathrm{PU}}_{c,w}(f; S) - R^{\mathrm{PU}}_{c,w}(f) \right| \le \frac{\varepsilon}{4} \right\}.$$

Bounding each source's contribution to the residual estimation error by $\varepsilon/8$, Hoeffding's inequality and a union bound give

$$\Pr(\mathcal{E}^c_{T_\varepsilon}) \le 2N_{T_\varepsilon}|\mathcal{C}||\mathcal{W}| \left[\exp\left(-\frac{n_+\varepsilon^2}{128\pi_+^2} \right) + \exp\left(-\frac{n_u\varepsilon^2}{128} \right) \right] \le \delta.$$

In particular, when $n_+ = n_u = n$,

$$\Pr(\mathcal{E}^c_{T_\varepsilon}) \le 4N_{T_\varepsilon}|\mathcal{C}||\mathcal{W}| \exp\left(-\frac{n\varepsilon^2}{128} \right).$$

This is one event for the fixed samples, and it controls the sample-dependent iterates through their membership in $\mathcal{F}_{\le T_\varepsilon}$.

Work on $\mathcal{E}_{T_\varepsilon}$. Whenever an update is selected at $t \le T_\varepsilon$, the empirical threshold implies

$$\tau_t R^{\mathrm{PU}}_{c_t,w_t}(f_t) \ge \tau_t \widehat{R}^{\mathrm{PU}}_{c_t,w_t}(f_t; S) - \frac{\varepsilon}{4} > \frac{\varepsilon}{2}.$$

Set $\Psi(f) = \mathbb{E}[(f(X) - f^\star(X))^2]$ and $d_t(x) = \tau_t c_t(x) w_t(f_t(x))$. Since $|d_t| \le 1$, $f^\star(x) \in [0, 1]$, and projection onto $[0, 1]$ cannot increase the distance to $f^\star(x)$, expanding the square gives

$$\begin{aligned}
\Psi(f_{t+1}) &\le \mathbb{E}\left[\left(f_t(X) + \frac{\varepsilon}{2} d_t(X) - f^\star(X) \right)^2 \right] \\
&= \Psi(f_t) - \varepsilon\tau_t R^{\mathrm{PU}}_{c_t,w_t}(f_t) + \frac{\varepsilon^2}{4}\mathbb{E}[d_t(X)^2] \\
&\le \Psi(f_t) - \frac{\varepsilon^2}{4}.
\end{aligned}$$

Here we used $R^{\mathrm{PU}}_{c,w}(f) = \mathbb{E}[c(X)w(f(X))(f^\star(X) - f(X))]$.

If the procedure performed $T_\varepsilon + 1$ successful updates, the preceding inequality would apply at $t = 0, \ldots, T_\varepsilon$, yielding

$$0 \le \Psi(f_{T_\varepsilon+1}) \le D_0 - \frac{(T_\varepsilon + 1)\varepsilon^2}{4} < 0,$$

a contradiction. Thus the empirical stopping condition is met after $\widehat{T} \le T_\varepsilon$ updates. At termination, $\sup_{c,w} |\widehat{R}^{\mathrm{PU}}_{c,w}(f_{\widehat{T}}; S)| \le 3\varepsilon/4$, and $f_{\widehat{T}} \in \mathcal{F}_{\le T_\varepsilon}$. The same event $\mathcal{E}_{T_\varepsilon}$ therefore gives

$$\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f_{\widehat{T}}) = \sup_{c,w} |R^{\mathrm{PU}}_{c,w}(f_{\widehat{T}})| \le \frac{3\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.$$

In particular, the stopping and multicalibration conclusions are simultaneous:

$$\Pr\left(\bigcup_{t=0}^{T_\varepsilon} \{\widehat{T} = t,\ \mathrm{MC}_{\mathcal{C},\mathcal{W}}(f_t) \le \varepsilon\} \right) \ge 1 - \delta.$$

Expanding $\log(2K^{T_\varepsilon+2}/\delta) = (T_\varepsilon + 2)\log K + \log(2/\delta)$ gives the stated sufficient total budget. The samples are drawn only once, so $N_+ = n_+$ and $N_u = n_u$. □

Neither D_0 nor T_ε is an input to Algorithm 1. Since $D_0 \le 1$, one may replace T_ε by $\lceil 4/\varepsilon^2 \rceil$ when choosing the sample sizes. If $f_0, \mathcal{C}, \mathcal{W}$ are obtained from independent training data, the argument applies conditionally on them.

VC classes and Cutoff calibration. The finite-class assumption can be relaxed when the witness classes have finitely many evaluation patterns on every finite sample. For a class $\mathcal{A}$, let $\Pi_{\mathcal{A}}(m)$ be the maximum number of distinct evaluation vectors on m points in its domain. A finite class satisfies $\Pi_{\mathcal{A}}(m) \le |\mathcal{A}|$; an indicator class of VC dimension $d_{\mathcal{A}}$ satisfies $\Pi_{\mathcal{A}}(m) \le (m+1)^{d_{\mathcal{A}}}$ by Sauer's lemma [34].

Fix m points and write $q_m = 2\Pi_{\mathcal{C}}(m)\Pi_{\mathcal{W}}(m)$. For each evaluation vector of f, an update has at most q_m possible evaluation vectors: the sign, the evaluations of c, and the evaluations of w at the current scores determine the next vector. Consequently, each of the two source classes

$$\mathcal{G}_+ = \{c\, w(f) : f \in \mathcal{F}_{\le T_\varepsilon},\ c \in \mathcal{C},\ w \in \mathcal{W}\},$$
$$\mathcal{G}_u = \{c\, w(f) f : f \in \mathcal{F}_{\le T_\varepsilon},\ c \in \mathcal{C},\ w \in \mathcal{W}\}$$

has at most

$$M_{T_\varepsilon}(m) \le \Pi_{\mathcal{C}}(m)\Pi_{\mathcal{W}}(m) \sum_{t=0}^{T_\varepsilon} q_m^t \le q_m^{T_\varepsilon+2}$$

distinct evaluation vectors on these points. Multiplication by f in $\mathcal{G}_u$ introduces no additional choices once its evaluation vector is fixed.

Apply symmetrization and the finite-vector Rademacher bound to these classes, followed by bounded-differences concentration [1, 34]. For $n_+ = n_u = n$, with probability at least $1-\delta$,

$$\sup_{f,c,w} |\widehat{R}^{\mathrm{PU}}_{c,w}(f;S) - R^{\mathrm{PU}}_{c,w}(f)| \le C_0(1+\pi_+)\sqrt{\frac{(T_\varepsilon+2)\log\big(2\Pi_{\mathcal{C}}(n)\Pi_{\mathcal{W}}(n)\big) + \log(4/\delta)}{n}},$$

where the supremum ranges over $f \in \mathcal{F}_{\le T_\varepsilon}$, $c \in \mathcal{C}$, and $w \in \mathcal{W}$, and $C_0 > 0$ is universal. Choosing this bound to be at most $\varepsilon/4$ gives the same potential-decrease and stopping argument. In particular, for indicator classes with finite VC dimensions $d_{\mathcal{C}}, d_{\mathcal{W}}$, it suffices to take

$$n \ge \frac{C}{\varepsilon^2}\left\{(T_\varepsilon+2)\big[1 + (d_{\mathcal{C}} + d_{\mathcal{W}})\log(n+1)\big] + \log(4/\delta)\right\}.$$

where $C > 0$ is universal. Since $T_\varepsilon \le \lceil 4/\varepsilon^2 \rceil$, this yields a sufficient total sample budget per source of $\widetilde{O}((1 + d_{\mathcal{C}} + d_{\mathcal{W}})/\varepsilon^4 + \log(1/\delta)/\varepsilon^2)$.

For example, $\mathcal{W} = \{v \mapsto \mathbf{1}\{v \in I\} : I \subseteq [0,1] \text{ is an interval}\}$ has VC dimension two. With $\mathcal{C} = \{1\}$, the corresponding criterion is Cutoff Calibration Error [37]; a subgroup class gives its multicalibration counterpart. This also covers any finite bounded real-valued $\mathcal{C}$, using $\Pi_{\mathcal{C}}(n) \le |\mathcal{C}|$, together with all interval indicators. General infinite real-valued classes require separate control of the reachable residual classes.

E.3 Other weak-supervision models

The same argument as in the finite-class PU case yields the promised $O(\varepsilon^{-4})$ total weak-sample order for the other weak-supervision examples using fixed calibration samples. The only model-specific ingredient is the uniform estimation error of the corrected residuals.

Assume the same bounds and sample independence of $f_0, \mathcal{C}, \mathcal{W}$ as in Theorem 4, with finite $\mathcal{C}, \mathcal{W}$. Suppose a weak-supervision model has a nonempty finite source index set J, source laws Q_j, and positive envelope constants B_j. Draw independent calibration samples $S_j = \{X_i^j\}_{i=1}^{m_j} \sim Q_j^{m_j}$ once and reuse them. Write the corrected population and empirical moments as

$$R^{\mathrm{WSL}}_{c,w}(f) = \sum_{j\in J} \mathbb{E}_{Q_j}[g_{j,f,c,w}(X)], \qquad \widehat{R}^{\mathrm{WSL}}_{c,w}(f;S) = \sum_{j\in J} \frac{1}{m_j}\sum_{i=1}^{m_j} g_{j,f,c,w}(X_i^j),$$

where $|g_{j,f,c,w}(x)| \le B_j$ uniformly over f, c, w. The correction uses the exact parameters of the weak-observation model, so that

$$R^{\mathrm{WSL}}_{c,w}(f) = \mathbb{E}[c(X)w(f(X))(Y - f(X))] = \mathbb{E}[c(X)w(f(X))(f^\star(X) - f(X))].$$

Use the same reachable-class construction with

$$T_\varepsilon = \lceil 4/\varepsilon^2 \rceil, \qquad K = 2|\mathcal{C}||\mathcal{W}|, \qquad N_{T_\varepsilon} := |\mathcal{F}_{\le T_\varepsilon}| \le K^{T_\varepsilon+1}.$$

If

$$\sum_{j\in J} B_j \sqrt{\frac{2\log(|J|K^{T_\varepsilon+2}/\delta)}{m_j}} \le \frac{\varepsilon}{4},$$

then, with probability at least $1-\delta$, Algorithm 1 with $\lambda = \varepsilon/2$ stops after at most T_ε successful updates and returns a $(\mathcal{C}, \mathcal{W}, \varepsilon)$-multicalibrated predictor. A sufficient source-wise condition is

$$m_j \ge \frac{32|J|^2 B_j^2}{\varepsilon^2} \log \frac{|J|K^{T_\varepsilon+2}}{\delta}, \qquad j \in J.$$

Here m_j is the total number of observations from source Q_j.

For UU, $J = \{1, 2\}$, $Q_j = P_{U_j}$, and $g_{j,f,c,w}(x) = \alpha_j(f(x))c(x)w(f(x))$. Thus $B_j = A_j$, with $A_j = \sup_{v\in[0,1]} |\alpha_j(v)|$ as in Proposition 2. The class prior and mixture proportions are known, and $\Delta \neq 0$. For Pconf, use the exact-confidence observation model of Corollary 3, including $P_- \ll P_+$, known π_+, and $r(X) = P(Y = 1 \mid X)$. Then $J = \{+\}$, $Q_+ = P_+$, and

$$g_{+,f,c,w}(x) = \pi_+ c(x)w(f(x))(1 - f(x)/r(x)).$$

Under $r(x) \ge C_r > 0$ almost surely under P_+, the existing envelope $B_+ = \pi_+(1 + 1/C_r)$ is valid.

Proof. For every fixed (f, c, w) and source j, Hoeffding's inequality gives

$$\Pr\left(\left|\frac{1}{m_j}\sum_{i=1}^{m_j} g_{j,f,c,w}(X_i^j) - \mathbb{E}_{Q_j}[g_{j,f,c,w}(X)]\right| > B_j\sqrt{\frac{2\log(|J|K^{T_\varepsilon+2}/\delta)}{m_j}}\right) \le \frac{2\delta}{|J|K^{T_\varepsilon+2}}.$$

A union bound over $J \times \mathcal{F}_{\le T_\varepsilon} \times \mathcal{C} \times \mathcal{W}$, which has cardinality $|J|N_{T_\varepsilon}|\mathcal{C}||\mathcal{W}| \le |J|K^{T_\varepsilon+2}/2$, and summation over sources show that

$$\Pr(\mathcal{E}_{\mathrm{WSL}}) \ge 1-\delta, \qquad \mathcal{E}_{\mathrm{WSL}} := \left\{ \sup_{\substack{f\in\mathcal{F}_{\le T_\varepsilon}\\ c\in\mathcal{C},\ w\in\mathcal{W}}} \left|\widehat{R}^{\mathrm{WSL}}_{c,w}(f;S) - R^{\mathrm{WSL}}_{c,w}(f)\right| \le \frac{\varepsilon}{4} \right\}.$$

The reachable class contains every realized iterate up to the deterministic horizon, so this event also controls their deviations.

On $\mathcal{E}_{\mathrm{WSL}}$, any selected update at $t \le T_\varepsilon$ satisfies

$$\tau_t R^{\mathrm{WSL}}_{c_t,w_t}(f_t) \ge \tau_t \widehat{R}^{\mathrm{WSL}}_{c_t,w_t}(f_t;S) - \frac{\varepsilon}{4} > \frac{\varepsilon}{2}.$$

The exact weak-residual rewrite therefore gives, by the same squared-potential calculation as in the PU proof,

$$\begin{aligned}\Psi(f_{t+1}) &\le \Psi(f_t) - \varepsilon\tau_t R^{\mathrm{WSL}}_{c_t,w_t}(f_t) + \frac{\varepsilon^2}{4}\mathbb{E}[(c_t(X)w_t(f_t(X)))^2] \\ &\le \Psi(f_t) - \frac{\varepsilon^2}{4}.\end{aligned}$$

If $T_\varepsilon + 1$ updates were performed, then

$$0 \le \Psi(f_{T_\varepsilon+1}) \le \Psi(f_0) - \frac{(T_\varepsilon+1)\varepsilon^2}{4} < 0,$$

because $\Psi(f_0) \le 1$, a contradiction. Thus the procedure stops with $\widehat{T} \le T_\varepsilon$. At termination, the empirical supremum is at most $3\varepsilon/4$, and the same event gives

$$\mathrm{MC}_{\mathcal{C},\mathcal{W}}(f_{\widehat{T}}) = \sup_{c,w} |R^{\mathrm{WSL}}_{c,w}(f_{\widehat{T}})| \le \sup_{c,w} |\widehat{R}^{\mathrm{WSL}}_{c,w}(f_{\widehat{T}};S)| + \frac{\varepsilon}{4} \le \varepsilon.$$

Finally, $\log(|J|K^{T_\varepsilon+2}/\delta) = (T_\varepsilon+2)\log K + \log(|J|/\delta)$. For the fixed source count in UU and Pconf, a sufficient total budget is

$$N_j = m_j = \mathcal{O}\left(B_j^2\left[\frac{\log(2|\mathcal{C}||\mathcal{W}|)}{\varepsilon^4} + \frac{\log(1/\delta)}{\varepsilon^2}\right]\right).$$

□

Thus the population boosting decrease is identical after the weak residual rewrite, and the model-specific envelopes determine how many fixed calibration samples suffice to control every reachable predictor.

E.4 Corrected weak objectives for Platt and temperature scaling

Let g_θ denote either g_β or $g_{a,b}$, and write

$$\ell_\theta^+(x) := -\log g_\theta(x), \qquad \ell_\theta^-(x) := -\log(1 - g_\theta(x)).$$

The fully supervised calibration objective is $\mathbb{E}[\ell_\theta^+(X)\mathbf{1}\{Y=1\} + \ell_\theta^-(X)\mathbf{1}\{Y=0\}]$. The weak baselines in Section 5.2 minimize corrected or stabilized empirical versions of this objective. For PU, the unbiased corrected objective is

$$\widehat{L}_{\mathrm{PU}}(\theta) = \frac{\pi_+}{n_+}\sum_{i=1}^{n_+}\{\ell_\theta^+(X_i^+) - \ell_\theta^-(X_i^+)\} + \frac{1}{n_u}\sum_{j=1}^{n_u}\ell_\theta^-(X_j^u).$$

The supplied implementation uses the non-negative PU stabilization of Kiryo et al. [28] for PU Platt and temperature scaling on both tabular and large-model data, optimizing

$$\widehat{L}_{\mathrm{nnPU}}(\theta) = \frac{\pi_+}{n_+}\sum_{i=1}^{n_+}\ell_\theta^+(X_i^+) + \max\left\{0, \frac{1}{n_u}\sum_{j=1}^{n_u}\ell_\theta^-(X_j^u) - \frac{\pi_+}{n_+}\sum_{i=1}^{n_+}\ell_\theta^-(X_i^+)\right\}.$$

This prevents the empirical negative-risk component from becoming negative and mitigates finite-sample overfitting. The stabilization is used only to fit the parametric PU scaling map; PU MC estimation and WLMC auditing use the corrected residual moments. For UU/MCD with $P_{U_k} = \theta_k P_+ + (1-\theta_k)P_-$ and $\Delta = \theta_1 - \theta_2$, the objective is obtained by substituting the log-loss vector into M_{UU}^{-1}:

$$\widehat{L}_{\mathrm{UU}}(\theta) = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(1-\theta_2)\pi_+\ell_\theta^+(X_i^{U_1}) - \theta_2\pi_-\ell_\theta^-(X_i^{U_1})}{\Delta} + \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{-(1-\theta_1)\pi_+\ell_\theta^+(X_j^{U_2}) + \theta_1\pi_-\ell_\theta^-(X_j^{U_2})}{\Delta}.$$

The UU implementation uses this objective without non-negative clipping; the non-negative UU variant combines the two sources within each class-specific risk component, clips each component below at zero, and then adds them.

For Pconf, Bayes' rule gives the corrected positive-sample objective

$$\widehat{L}_{\mathrm{Pconf}}(\theta) = \frac{\pi_+}{n}\sum_{i=1}^{n}\left\{\ell_\theta^+(X_i) + \frac{1-r_i}{r_i}\ell_\theta^-(X_i)\right\}.$$

In experiments, these objectives are optimized over the one-dimensional temperature parameter or the two-dimensional Platt parameters on a held-out weak calibration split, keeping the base predictor fixed. For each fixed post-processing parameter, the unmodified corrected objectives follow from the WSL decontamination identities and source-wise empirical averaging. Non-negative risk correction is an optimization stabilization and is not an unbiased risk rewrite.

F Settings of Toy Data Experiments

The toy experiment in Figure 1 uses a one-dimensional binary classification problem with a known data-generating distribution. We draw $X \sim \mathrm{Unif}([0,1])$ and evaluate a fixed probabilistic predictor

$$f(x) = \mathrm{clip}(0.12 + 0.76x, 10^{-6}, 1 - 10^{-6}).$$

The true conditional label probability is

$$r(x) = \Pr(Y=1 \mid X=x) = \mathrm{clip}\{f(x) + 0.11\sin(2\pi x), 0.02, 0.98\},$$

and labels follow $Y \mid X = x \sim \mathrm{Bernoulli}(r(x))$. The prevalence is $\pi_+ = \mathbb{E}[r(X)]$. The evaluated finite class is formed by crossing eight equal-width subgroups $\mathcal{C} = \{G_1, \ldots, G_8\}$ partitioning $\mathcal{X}$ $([0,1])$ with ten equal-width score bins $\mathcal{W} = \{B_1, \ldots, B_{10}\}$ on $[0,1]$. Thus each witness cell is

$$a_{g,b}(x; f) := \mathbf{1}\{x \in G_g\}\mathbf{1}\{f(x) \in B_b\}.$$

For this diagnostic, the population target is the finite-class multicalibration moment

$$\begin{aligned}\mathrm{MC}_{\mathrm{toy}}(f) &:= \max_{g\in[8],\, b\in[10]} \left|\mathbb{E}\left[a_{g,b}(X;f)(Y-f(X))\right]\right| \\ &= \max_{g\in[8],\, b\in[10]} \left|\mathbb{E}\left[a_{g,b}(X;f)(r(X)-f(X))\right]\right|.\end{aligned}$$

This is the unnormalized residual moment used in the multicalibration definition; it is not divided by the mass of the corresponding cell. The population value is computed numerically using a grid of 400,000 points over $[0,1]$. The plotted sample sizes are $128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768$, and 65536.

For each sample size, independent Monte Carlo repetitions are used to construct PN, Pconf, PU, and UU estimates of the same target. The PN estimate uses labeled marginal samples (X_i, Y_i) and the empirical cell moments $n^{-1}\sum_i a_{g,b}(X_i;f)(Y_i - f(X_i))$. The Pconf estimate uses positive samples $X_i^+ \sim P(X \mid Y=1)$ with the exact confidence $r(X_i^+)$ and the corrected moments

$$\frac{\pi_+}{n}\sum_{i=1}^{n} a_{g,b}(X_i^+;f)\left(1-\frac{f(X_i^+)}{r(X_i^+)}\right).$$

The PU estimate uses positive samples and marginal unlabeled samples $X_j^u \sim P_X$:

$$\frac{\pi_+}{n}\sum_{i=1}^{n} a_{g,b}(X_i^+;f) - \frac{1}{n}\sum_{j=1}^{n} a_{g,b}(X_j^u;f)f(X_j^u).$$

The UU estimate uses two unlabeled mixtures $P_{U_k} = \rho_k P_+ + (1-\rho_k)P_-$ with $\rho_1 = 0.80$ and $\rho_2 = 0.20$, where $P_+ = P(X \mid Y=1)$ and $P_- = P(X \mid Y=0)$, and applies the corresponding two-source decontamination weights to estimate the same cell moments.

For a method $q \in \{\mathrm{PN}, \mathrm{Pconf}, \mathrm{PU}, \mathrm{UU}\}$, sample size n, and repetition s, let $\widehat{\mathrm{MC}}_n^{q,(s)}$ be the corresponding estimate. The plotted point is the mean absolute estimation error

$$\frac{1}{R}\sum_{s=1}^{R}\left|\widehat{\mathrm{MC}}_n^{q,(s)} - \mathrm{MC}_{\mathrm{toy}}(f)\right|,$$

where R is the number of Monte Carlo repetitions at that sample size; in the plotted experiment, $R = 10$. The vertical error bar is one standard deviation across those absolute errors,

$$\left[\frac{1}{R}\sum_{s=1}^{R}\left(\left|\widehat{\mathrm{MC}}_n^{q,(s)} - \mathrm{MC}_{\mathrm{toy}}(f)\right| - \frac{1}{R}\sum_{t=1}^{R}\left|\widehat{\mathrm{MC}}_n^{q,(t)} - \mathrm{MC}_{\mathrm{toy}}(f)\right|\right)^2\right]^{1/2}.$$

Thus the error bars show variability across Monte Carlo repetitions; they are not confidence intervals or standard errors. The gray guide lines in the plot have slope $-1/2$ on the log–log scale and indicate the parametric finite-sample rate predicted by the concentration bounds.

G Dataset and Subgroup Descriptions

This appendix describes the real-data evaluation setting used in Section 6. We follow the finite-subgroup viewpoint common in practical multicalibration experiments: rather than searching over an abstract infinite class, we evaluate a fixed collection of interpretable groups defined by dataset metadata or tabular attributes. The groups may overlap, so a point can belong to several evaluated subpopulations.

We first define the evaluation metrics, subgroup collections, and data splits. Appendix H then gives the authoritative split-first construction of the weak observation views used by those metrics. Later appendices summarize the model and post-processing hyperparameters, reproducibility details, and detailed result tables.

G.1 Calibration and multicalibration metrics

MC is the unnormalized witness moment appearing in the theory, so small subgroup–bin cells can have small absolute moments even when their conditional calibration error is large. To compare subpopulations fairly, we also report MaxECE: it normalizes each subgroup's binned calibration error by the subgroup mass and then takes the worst subgroup. Following Hansen et al. [21], our numerical experiments report ECE and MaxECE in addition to MC. Let $f : \mathcal{X} \to [0, 1]$ be a frozen predictor, let $B_1, \ldots, B_K$ be score bins, and let $G_1, \ldots, G_m$ be the evaluated subgroup collection. We also write $G_0 = \mathcal{X}$ for the whole population and

$$a_{g,b}(x; f) := \mathbf{1}\{x \in G_g\}\mathbf{1}\{f(x) \in B_b\}.$$

The superscript or denotes oracle clean-label quantities; in the fully supervised PN setting, these coincide with the corresponding full-label estimates. For a group–bin cell, define the following moment and group mass

$$m^{\mathrm{or}}_{g,b}(f) := \mathbb{E}\big[a_{g,b}(X; f)(Y - f(X))\big], \qquad \mu_g := \Pr(X \in G_g).$$

Given a clean sample $S = \{(X_i, Y_i)\}_{i=1}^n$, the corresponding oracle empirical quantities are

$$\widehat{m}^{\mathrm{or}}_{g,b}(f) := \frac{1}{n}\sum_{i=1}^n a_{g,b}(X_i; f)(Y_i - f(X_i)), \qquad \widehat{\mu}^{\mathrm{or}}_g := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in G_g\}.$$

The oracle ECE, maxECE, and MC columns in the tables are computed as

$$\widehat{\mathrm{ECE}}^{\mathrm{or}}(f) := \sum_{b=1}^K \left|\widehat{m}^{\mathrm{or}}_{0,b}(f)\right|, \tag{41}$$

$$\widehat{\mathrm{maxECE}}^{\mathrm{or}}(f) := \max_{g\in[m]:\, \widehat{\mu}^{\mathrm{or}}_g > \mu_{\min}} \frac{1}{\widehat{\mu}^{\mathrm{or}}_g}\sum_{b=1}^K \left|\widehat{m}^{\mathrm{or}}_{g,b}(f)\right|,$$

$$\widehat{\mathrm{MC}}^{\mathrm{or}}(f) := \max_{g\in[m],\, b\in[K]} \left|\widehat{m}^{\mathrm{or}}_{g,b}(f)\right|.$$

Here $\mu_{\min}$ is the minimum active-group mass used by the implementation; setting $\mu_{\min} = 0$ includes all groups with positive empirical mass. The ECE formula is the usual binned ECE because $\left|\widehat{m}^{\mathrm{or}}_{0,b}\right|$ equals the bin mass times the absolute difference between the average label and average prediction in bin B_b. The maxECE formula applies the same binned ECE calculation conditionally within each subgroup and then takes the worst subgroup. MC instead keeps the supremum-style unnormalized group–bin residual used by the witness definition.

Weak estimates of ECE, maxECE, and MC. Appendix H specifies how the held-out validation and test splits are converted into Pconf, PU, and UU views. Here we only define the corrected group–bin residuals that are plugged into the metrics and into the WLMC audit. For weak-observation model $q \in \{\mathrm{PU}, \mathrm{UU}, \mathrm{Pconf}\}$, define $\widehat{m}^q_{g,b}(f)$ by applying the corrected witness estimator to $c(x) = \mathbf{1}\{x \in G_g\}$ and $w(v) = \mathbf{1}\{v \in B_b\}$. Thus, for PU data (S_+, S_u),

$$\widehat{m}^{\mathrm{PU}}_{g,b}(f) = \frac{\pi_+}{n_+}\sum_{i=1}^{n_+} a_{g,b}(X^+_i; f) - \frac{1}{n_u}\sum_{j=1}^{n_u} a_{g,b}(X^u_j; f)f(X^u_j). \tag{42}$$

For UU data with $U_k = \rho_k P_+ + (1 - \rho_k)P_-$ and $\Delta = \rho_1 - \rho_2 \neq 0$,

$$\widehat{m}^{\mathrm{UU}}_{g,b}(f) = \sum_{k=1}^2 \frac{1}{n_k}\sum_{i=1}^{n_k} \alpha_k(f(X^{U_k}_i))a_{g,b}(X^{U_k}_i; f),$$

where

$$\alpha_1(v) = \frac{(1-\rho_2)\pi_+(1-v) + \rho_2\pi_- v}{\Delta}, \qquad \alpha_2(v) = \frac{-(1-\rho_1)\pi_+(1-v) - \rho_1\pi_- v}{\Delta}.$$

For Pconf data $\{(X_i, r_i)\}_{i=1}^n$ with $X_i \sim P_+$ and $r_i \approx \Pr(Y = 1 \mid X_i)$,

$$\widehat{m}^{\mathrm{Pconf}}_{g,b}(f) = \frac{\pi_+}{n}\sum_{i=1}^n a_{g,b}(X_i; f)\left(1 - \frac{f(X_i)}{r_i}\right), \tag{43}$$

Table 1: Real-data benchmark summary.

Dataset	Prediction target	Total n	Groups m	Subgroup description
ACSIncome [8]	income $>$ 50K	195,665	10	Groups based on demographic ACS attributes such as race, sex, age, marital status, and simple intersections.
CreditDefault [41]	default on credit-card debt	30,000	15	Groups based on sex, education, marital status, age buckets, and related intersections after preprocessing.
HMDA [6]	mortgage approval	114,185	13	Groups based on applicant and co-applicant ethnicity, sex, and race in the processed HMDA loan-application data.
MEPS [39]	high healthcare utilization	11,079	14	Groups based on age, race, region, poverty-related categories, and related health-survey metadata.
CelebA	`Blond_Hair`	202,599	18	Binary facial-attribute groups derived from CelebA metadata, including sex, age, hair, skin, and face-related attributes.
CivilComments	toxicity	447,998	13	Identity-mention groups for toxic-comment prediction, including gender, LGBTQ, religion, and race-related groups and their complements.

with r_i replaced by $\max\{r_i, \tau\}$ if confidence clipping is enabled.

The WLMC audit uses exactly this corrected residual table: it evaluates the finite subgroup–bin family, finds the largest corrected signed residual, and applies the corresponding post-processing update. Appendix H supplements this definition by specifying which split-specific source indices, confidence values, and subgroup masks enter the Pconf, PU, and UU versions of the table.

The weak ECE, maxECE, and MC columns use the same aggregation rules as Eq. (41), replacing oracle cell moments by the corrected weak moments above. For group-wise maxECE, the denominator is the target covariate group mass computed from the full feature-only held-out evaluation split, denoted $\widehat{\mu}_g^{\mathrm{eval}}$ and defined in Eq. (45). Hence, for $q \in \{\mathrm{PU}, \mathrm{UU}, \mathrm{Pconf}\}$,

$$\widehat{\mathrm{ECE}}^q(f) := \sum_{b=1}^{K} \left| \widehat{m}_{0,b}^q(f) \right|, \tag{44}$$

$$\widehat{\mathrm{maxECE}}^q(f) := \max_{g \in [m]:\, \widehat{\mu}_g^{\mathrm{eval}} > \mu_{\min}} \frac{1}{\widehat{\mu}_g^{\mathrm{eval}}} \sum_{b=1}^{K} \left| \widehat{m}_{g,b}^q(f) \right|,$$

$$\widehat{\mathrm{MC}}^q(f) := \max_{g \in [m],\, b \in [K]} \left| \widehat{m}_{g,b}^q(f) \right|.$$

This section therefore fixes the metric definitions; Appendix H is the source of truth for the split-first weak-data construction, the unbiasedness of $\widehat{\mu}_g^{\mathrm{eval}}$ in our semi-synthetic experiments, and the alternative Pconf/PU/UU denominators needed when no full target covariate evaluation pool is available. Oracle estimates are used only for reporting and analysis, while weak estimates are used for weak validation and weak post-processing selection.

G.2 Real-data datasets and subgroup collections

Table 1 summarizes the real-data benchmark setting. For the finite subgroup family, we adopt the dataset-specific subgroup collections used by the clean-label multicalibration benchmark of Hansen et al. [21]; the subgroup family is not tuned separately for our weak-supervision experiments. Each loader defines a finite, interpretable collection of protected-attribute groups, identity-mention groups, facial-attribute groups, and simple intersections. The same code-level group indicators are used for oracle metrics, weak estimates, validation-based method selection, WLMC audits, and the CDF/scatter diagnostics in Appendix L.4. All MC evaluations use 10 score bins, so the number of evaluated group–bin cells is $10m$, where m is the number of groups reported in Table 1. Table 1 gives short descriptions of the loader-defined subgroup collections; the exact code-level definitions are provided in the reproducibility code.

G.3 Data splits and effective sample sizes

The test split is created first with seed 42. The validation split is then created from the remaining data using the run seed. Post-processing uses a separate correction split carved out of the training partition

Table 2: Effective split sizes for the real-data experiments. Values are rounded because the exact counts depend on integer rounding in the split routine.

Dataset	base train	post-processing correction	validation	test
ACSIncome	$\approx 70{,}400$	$\approx 47{,}000$	$\approx 39{,}100$	$\approx 39{,}100$
CreditDefault	$\approx 10{,}800$	$\approx 7{,}200$	$\approx 6{,}000$	$\approx 6{,}000$
HMDA	$\approx 41{,}100$	$\approx 27{,}400$	$\approx 22{,}800$	$\approx 22{,}800$
MEPS	$\approx 4{,}000$	$\approx 2{,}700$	$\approx 2{,}200$	$\approx 2{,}200$
CelebA	$\approx 97{,}200$	$\approx 24{,}300$	$\approx 40{,}500$	$\approx 40{,}500$
CivilComments	$\approx 215{,}000$	$\approx 53{,}800$	$\approx 89{,}600$	$\approx 89{,}600$

with seed 50. The tabular data uses correction fraction 0.4 within the training partition, giving roughly $0.36n$ base-training examples and $0.24n$ correction examples. CelebA and CivilComments use correction fraction 0.2, giving roughly $0.48n$ base-training examples and $0.12n$ correction examples. In all cases, validation and test use roughly $0.20n$ each. These split sizes refer to the underlying fully labeled data before labels are hidden; Appendix H explains how PN, Pconf, PU, and UU views are generated separately within each split.

Tabular preprocessing. The tabular preprocessing follows the repository dataloaders. ACSIncome uses the folktables ACSIncome setup with the binary income label. CreditDefault drops the ID column, bins several continuous financial variables into terciles, and one-hot encodes categorical columns. HMDA uses the processed Texas loan-application files, maps action codes to a binary approval label, drops denial-reason and constant columns, and one-hot encodes low-cardinality categorical variables. MEPS uses the provided processed feature/label files with cleaned column names. Standardization, when used by a model, is fit only on the training split.

G.4 Random seeds and reported variability

The toy convergence plot and result tables summarize repeated runs, whereas the tabular scatter plots display individual runs. The appendix reports variability for aggregated results. The toy convergence experiments use 10 random seeds. The tabular real-data experiments use 10 random seeds, varying the train/validation split and the randomized components of the base predictor when applicable. The CelebA and CivilComments large-model experiments use 5 random seeds. Oracle test metrics are used for final reporting and descriptive comparisons, not for fitting or validation-based model selection.

H Data preparation under weak supervision

This appendix is the source of truth for the experimental weak-data construction. Appendix G defines the finite subgroup–bin metrics and the corrected weak MC residuals used by WLMC; the present appendix complements that definition by specifying exactly when the data are split, how Pconf, PU, and UU views are generated inside each split, and why the maxECE denominator used in the experiments estimates the target covariate marginal.

All weak views in the real-data experiments are semi-synthetic views of the same underlying labeled data. This protocol is important for two reasons. First, it lets us evaluate weak-observation methods against oracle clean-label metrics. Second, it makes the evaluation population invariant across PN, Pconf, PU, and UU: only the observation model changes, not the held-out covariate population on which calibration is evaluated.

Order of splitting and weak-view generation. For each dataset and random seed, we first split the fully labeled dataset into disjoint supervised splits. The training portion is further divided into a base-training split and a post-processing correction split; the remaining held-out data are used for validation and test evaluation. We write these four splits as

$$D_{\mathrm{tr}}, \qquad D_{\mathrm{cal}}, \qquad D_{\mathrm{val}}, \qquad D_{\mathrm{test}}.$$

The weak observations are generated *after* this split. In particular, for each $S \in \{D_{\mathrm{tr}}, D_{\mathrm{cal}}, D_{\mathrm{val}}, D_{\mathrm{test}}\}$, the Pconf, PU, and UU views are constructed separately using only examples inside that split. We never first convert the full dataset into a weak dataset and then split it. Thus,

the weak training view is generated from D_{tr}, the weak post-processing view from D_{cal}, the weak validation view from D_{val}, and the weak test view from D_{test}.

For a generic split $S = \{(X_i, Y_i)\}_{i \in I_S}$, let

$$I_S^+ := \{i \in I_S : Y_i = 1\}, \qquad I_S^- := \{i \in I_S : Y_i = 0\}, \qquad \widehat{\pi}_{+,S} := |I_S^+|/|I_S|.$$

The hidden labels in S are used by the experimenter only to simulate the weak observation source and to compute oracle reference metrics. The weak base learner, weak validation criterion, and weak post-processing algorithm receive only the corresponding weak view described below.

PN. The PN view observes the clean labeled pairs in the corresponding split,

$$D_S^{\mathrm{PN}} = \{(X_i, Y_i) : i \in I_S\}.$$

This is the usual fully supervised view. In the PN setting, the weak and oracle ECE formulas coincide when evaluated on the same scores and bins.

Pconf. The Pconf view keeps only positive examples from the split and attaches a positive-confidence value $r_i \approx \Pr(Y = 1 \mid X_i)$. Concretely,

$$D_S^{\mathrm{Pconf}} = \{(X_i, r_i) : i \in I_S^+\}.$$

In the implementation, r_i is obtained from a teacher score; the default setting uses this teacher score without additional perturbation. The class prior $\widehat{\pi}_{+,S}$ is also recorded for the Pconf correction. Thus, the Pconf learner or post-processor sees positive examples, confidence values, and the supplied prior, but not negative labels in that split. These teacher-based confidences are estimates, so the exact-confidence Pconf guarantee does not directly apply to the real-data runs; clipping does not remove confidence-estimation error.

PU. The PU view uses a case-control / SCAR two-sample construction inside the split. The positive bag is sampled from the positive pool I_S^+, while the unlabeled bag is sampled from the full split I_S. With default rates $\lambda_P = 0.5$ and $\lambda_U = 1.0$, the implementation draws

$$n_P = \mathrm{round}(\lambda_P |I_S^+|), \qquad n_U = \mathrm{round}(\lambda_U |I_S|)$$

indices. The positive bag is

$$S_P = \{X_i^+ : i = 1, \ldots, n_P\}, \qquad X_i^+ \sim P_+,$$

and the unlabeled bag is

$$S_U = \{X_j^u : j = 1, \ldots, n_U\}, \qquad X_j^u \sim P_X.$$

Semi-synthetically, these samples are drawn from I_S^+ and I_S, respectively. By default, both draws are with replacement and independent. Therefore the same underlying example may appear more than once, a positive example may also appear in the unlabeled bag, and the two bags should be viewed as empirical draws from two source distributions rather than as a partition of the split. The class prior $\widehat{\pi}_{+,S}$ is computed from the underlying split and used in the PU correction.

UU. The UU view constructs two unlabeled mixture sources inside the split. In code notation,

$$U_1 = (1 - \gamma_1)P_+ + \gamma_1 P_-, \qquad U_2 = \gamma_2 P_+ + (1 - \gamma_2)P_-.$$

The default values are $\gamma_1 = \gamma_2 = 0.2$, so the two sources have target positive fractions 0.8 and 0.2, respectively. With default source rates $\lambda_1 = \lambda_2 = 1.0$, the implementation draws

$$n_1 = \mathrm{round}(\lambda_1 |I_S|), \qquad n_2 = \mathrm{round}(\lambda_2 |I_S|)$$

examples for the two sources. Source 1 is formed by drawing approximately $\mathrm{round}((1 - \gamma_1)n_1)$ positive examples and the remaining examples from the negative pool; source 2 is formed by drawing approximately $\mathrm{round}(\gamma_2 n_2)$ positive examples and the remaining examples from the negative pool. The draws are with replacement and then shuffled. The two sources are independently generated from the same split, and $\widehat{\pi}_{+,S}$ is recorded for the UU decontamination coefficients.

Training, post-processing, and evaluation usage. The base predictor is trained on the view of D_{tr} specified by the experimental scenario: PN, Pconf, PU, or UU. Post-hoc recalibration is fit on the corresponding view of D_{cal}. Weak validation quantities are computed from the weak view of D_{val}, and final weak estimates are computed from the weak view of D_{test}. Oracle validation or test labels are used only for oracle reporting and analysis, not for weak post-processing.

Evaluation denominator for maxECE. This paragraph expands the denominator convention used in Eq. (44). For the reported weak ECE, maxECE, and MC estimates, the signed group–bin residual moment is estimated using the regime-specific weak correction, as in Eqs. (42)–(43). However, the group denominator in maxECE is computed from the invariant held-out covariate pool of the same evaluation split. For an evaluation split $E \in \{D_{\mathrm{val}}, D_{\mathrm{test}}\}$, define

$$E^X := \{X_i^{\mathrm{eval}} : i = 1, \ldots, n_{\mathrm{eval}}\}.$$

Because weak views are generated after the split, this covariate pool is a held-out sample from the target marginal $P_X = \pi_+ P_+ + (1 - \pi_+)P_-$, containing both $Y = 1$ and $Y = 0$ examples according to the target class mixture. Thus, for every group G_g, the implementation uses

$$\widehat{\mu}_g^{\mathrm{eval}} := \frac{1}{n_{\mathrm{eval}}} \sum_{i=1}^{n_{\mathrm{eval}}} \mathbf{1}\{X_i^{\mathrm{eval}} \in G_g\}. \tag{45}$$

This estimator uses no labels. Under the random split model, it is an unbiased estimator of the target group mass,

$$\mathbb{E}[\widehat{\mu}_g^{\mathrm{eval}}] = P_X(X \in G_g).$$

Consequently, in our semi-synthetic experimental protocol, the same denominator is valid for PN, Pconf, PU, and UU. The weak-observation correction is applied to the numerator residual moments, not to the definition of the evaluation population. In particular, for $q \in \{\mathrm{Pconf}, \mathrm{PU}, \mathrm{UU}\}$, the reported group-wise weak maxECE has the form

$$\widehat{\mathrm{maxECE}}^q(f) = \max_{g:\widehat{\mu}_g^{\mathrm{eval}} > \mu_{\min}} \frac{1}{\widehat{\mu}_g^{\mathrm{eval}}} \sum_{b=1}^{K} |\widehat{m}_{g,b}^q(f)|,$$

where $\widehat{m}_{g,b}^q(f)$ is the corrected weak cell residual. The denominator $\widehat{\mu}_g^{\mathrm{eval}}$ itself is unbiased for $P_X(G_g)$; the ratio defining maxECE is a plug-in finite-sample estimate of the corresponding normalized calibration quantity.

If no full evaluation covariate pool is available. The denominator choice above is specific to our controlled semi-synthetic experiments, where the full feature-only validation and test pools are retained even when labels are hidden from the weak estimator. Here we discuss the case where, in a genuinely weak-only evaluation setting, one may not have access to such an invariant held-out covariate pool. Then the group mass $P_X(G_g)$ must be estimated from the available weak source using the observation-model-specific marginal correction. For PU, the unlabeled bag is already drawn from P_X, so one may use

$$\widehat{\mu}_g^{\mathrm{PU}} = \frac{1}{n_U} \sum_{j=1}^{n_U} \mathbf{1}\{X_j^u \in G_g\}.$$

For Pconf, if only positive-confidence data (X_i^+, r_i) are available, then

$$P_X(G_g) = \pi_+ \mathbb{E}_{X \sim P_+}\left[\frac{\mathbf{1}\{X \in G_g\}}{r(X)}\right],$$

so the corresponding plug-in estimator is

$$\widehat{\mu}_g^{\mathrm{Pconf}} = \frac{\pi_+}{n_+} \sum_{i=1}^{n_+} \frac{\mathbf{1}\{X_i^+ \in G_g\}}{r_i},$$

with the same confidence clipping convention as in the residual estimator if clipping is used. For UU, let

$$\widehat{q}_{k,g} := \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{1}\{X_i^{U_k} \in G_g\}, \qquad k = 1, 2,$$

and let $\delta = 1 - \gamma_1 - \gamma_2 \neq 0$. Since $U_1 = (1-\gamma_1)P_+ + \gamma_1 P_-$ and $U_2 = \gamma_2 P_+ + (1-\gamma_2)P_-$, the target marginal group mass is estimated by

$$\widehat{\mu}_g^{\mathrm{UU}} = \frac{\pi_+ - \gamma_2}{\delta}\widehat{q}_{1,g} + \frac{(1-\pi_+) - \gamma_1}{\delta}\widehat{q}_{2,g}.$$

These alternative denominators are not used in our reported semi-synthetic experiments because Eq. (45) is available and directly estimates the target evaluation population. They would be the appropriate denominators in deployments where the only accessible evaluation data are the Pconf or UU weak sources themselves.

Validation and model selection. We distinguish two uses of validation. The first is base-predictor selection, such as choosing among fixed candidate predictors or selecting a neural model by validation performance. In the PN setting, the validation view is the clean labeled split, and the validation score is ordinary accuracy:

$$\widehat{\mathrm{Acc}}_{\mathrm{val}}(f) = \frac{1}{|I_{\mathrm{val}}|}\sum_{i \in I_{\mathrm{val}}} \mathbf{1}\{\mathbf{1}[f(X_i) \geq 1/2] = Y_i\}.$$

PN base candidates are selected by maximizing this quantity. This is the only case in which clean validation labels are used to select a base predictor.

For weakly trained base predictors, the clean labels in D_{val} are hidden from the selection rule. We instead construct the same weak view of D_{val} as in the corresponding training regime and select the candidate with the smallest weak validation loss. Let $s_f(x)$ denote the logit of candidate f, and define $\ell_+(s) = \log(1+e^{-s})$ and $\ell_-(s) = \log(1+e^{s})$. For Pconf, using validation positive-confidence examples (X_i^+, r_i), the criterion is

$$\widehat{L}_{\mathrm{val}}^{\mathrm{Pconf}}(f) = \frac{\widehat{\pi}_{+,\mathrm{val}}}{n_+}\sum_{i=1}^{n_+}\left\{\ell_+(s_f(X_i^+)) + \frac{1-r_i}{r_i}\ell_-(s_f(X_i^+))\right\},$$

with the same clipping convention for r_i as in the Pconf estimator. For PU, using the validation positive bag S_P^{val} and unlabeled bag S_U^{val}, the tabular validation criterion is the corrected PU validation risk

$$\widehat{L}_{\mathrm{val}}^{\mathrm{PU}}(f) = \frac{\widehat{\pi}_{+,\mathrm{val}}}{n_P}\sum_{i=1}^{n_P}\ell_+(s_f(X_i^+)) + \frac{1}{n_U}\sum_{j=1}^{n_U}\ell_-(s_f(X_j^u)) - \frac{\widehat{\pi}_{+,\mathrm{val}}}{n_P}\sum_{i=1}^{n_P}\ell_-(s_f(X_i^+)).$$

Thus the positive bag contributes the positive-risk term and the positive correction term, while the unlabeled bag estimates the marginal negative-loss term.

For UU, using the two validation mixture sources U_1^{val} and U_2^{val}, we use the contamination-matrix inverse. Let $\widehat{\mathbb{E}}_{1,\mathrm{val}}$ and $\widehat{\mathbb{E}}_{2,\mathrm{val}}$ denote empirical averages over these two sources and set $\delta = 1-\gamma_1-\gamma_2$. The corrected validation loss is

$$\begin{aligned}\widehat{L}_{\mathrm{val}}^{\mathrm{UU}}(f) = \widehat{\pi}_{+,\mathrm{val}}&\left\{\frac{1-\gamma_2}{\delta}\widehat{\mathbb{E}}_{1,\mathrm{val}}[\ell_+(s_f(X))] - \frac{\gamma_1}{\delta}\widehat{\mathbb{E}}_{2,\mathrm{val}}[\ell_+(s_f(X))]\right\}\\ &+ (1-\widehat{\pi}_{+,\mathrm{val}})\left\{-\frac{\gamma_2}{\delta}\widehat{\mathbb{E}}_{1,\mathrm{val}}[\ell_-(s_f(X))] + \frac{1-\gamma_1}{\delta}\widehat{\mathbb{E}}_{2,\mathrm{val}}[\ell_-(s_f(X))]\right\}.\end{aligned}$$

For the non-negative UU variant, the same two-source decomposition is used with the non-negative correction applied to the partial risks, matching the training objective. The hidden labels in D_{val} are used only by the experimenter to generate these weak validation views and to compute oracle reference diagnostics; they are not used by the weak selection rule.

The second use of validation is post-processing selection after the base predictor has been frozen. All post-processing candidates are fit on D_{cal}, not on validation or test. They are then selected on D_{val} by the supervision-appropriate calibration criterion for the reported table or figure. Clean-label PN selection computes this criterion from the PN validation labels, while weak selection computes the corresponding weak estimate from $D_{\mathrm{val}}^{\mathrm{Pconf}}$, $D_{\mathrm{val}}^{\mathrm{PU}}$, or $D_{\mathrm{val}}^{\mathrm{UU}}$. For MC-focused post-processing comparisons, for example, weak post-processing methods are selected by the weak validation estimate of MC violation, whereas PN/oracle diagnostics use the clean-label validation analogue. Accuracy is not used to select post-processing methods; it is reported on the held-out test split as auxiliary predictive-performance information. Test labels are not used until the final evaluation.

Use of nnPU and ordinary PU objectives. For reference, the non-negative PU principle of Kiryo et al. [28] clips the empirical negative-risk component when that stabilized objective is used. Let $s_\theta(x) \in \mathbb{R}$ denote the model logit. With logistic losses $\ell_+(s) = \log(1+e^{-s})$ and $\ell_-(s) = \log(1+e^{s})$, the empirical PU decomposition is

$$\widehat{R}_+^{\mathrm{PU}}(\theta) := \pi_+ \widehat{\mathbb{E}}_{P_+}[\ell_+(s_\theta(X))],$$
$$\widehat{R}_-^{\mathrm{PU}}(\theta) := \widehat{\mathbb{E}}_{P_X}[\ell_-(s_\theta(X))] - \pi_+ \widehat{\mathbb{E}}_{P_+}[\ell_-(s_\theta(X))].$$

The nnPU training objective clips the empirical negative-risk component,

$$\widehat{R}_{\mathrm{nnPU}}(\theta) = \widehat{R}_+^{\mathrm{PU}}(\theta) + \max\{0, \widehat{R}_-^{\mathrm{PU}}(\theta)\}. \tag{46}$$

The supplied implementation uses this clipped nnPU objective for PU base learning in CelebA and CivilComments and for PU Platt and temperature post-processing on both tabular and large-model data. Tabular base learning uses the corresponding supervised-loss or fitting-criterion rewrite. Weak MC/WLMC and calibration evaluation use the corrected residual estimators without applying this non-negative risk correction.

I Real-data model and post-processing hyperparameters

This section records the base predictors, fixed hyperparameters, and post-processing hyperparameters used in the real-data experiments. The model families and tabular hyperparameter settings follow the clean-label multicalibration benchmark of Hansen et al. [21]. Our weak-supervision experiments keep the same base architecture within each dataset and replace only the supervision source and objective as described in Appendix H. PN base models use oracle validation accuracy for base-model selection. Weakly trained base models use the weak validation objectives described in Appendix H; no clean validation labels are used to select weak candidates. All reported post-processing methods freeze the selected base predictor and are fit only on the correction split.

I.1 Tabular base predictors

The tabular data uses the same finite feature representations and subgroup metadata as in Appendix G. The reported tabular experiments use six base-predictor families: decision trees, random forests, logistic regression, support vector machines, naive Bayes, and MLP. We follow the same benchmark candidate grids used in the clean-label benchmark and use the selected settings for correction fraction 0.4; the weak experiments do not run a new hyperparameter search. For tabular weak baselines, the base predictor is obtained by minimizing the regime-specific empirical objective obtained from the WSL invariant-estimator rewrite of the corresponding supervised loss or fitting criterion.

For decision trees, the selected model controls tree depth and the minimum number of samples required to split an internal node. For random forests, we use 100-tree ensembles with the selected tree-depth and split-size controls for each dataset. The random-forest predictor is also used as the Pconf teacher in the tabular construction.

For logistic regression, we use the benchmark-selected amount of regularization. For support vector machines, we use the benchmark-selected margin-regularization setting with standardized features. Naive Bayes is treated as a hyperparameter-free baseline. Here MLP denotes a fully connected feed-forward binary classifier with ReLU hidden layers and two output logits, trained with cross-entropy and the Adam optimizer on standardized features. At correction fraction 0.4, the MLP configurations are as follows: ACSIncome uses architecture 10–128–BN–256–BN–128–2, batch size 64, learning rate 10^{-3}, no weight decay, and 50 epochs; CreditDefault uses 118–128–256–128–2, batch size 128, learning rate 10^{-2}, no weight decay, and 5 epochs; HMDA uses 89–100–2, batch size 128, learning rate 10^{-3}, weight decay 10^{-5}, and 30 epochs; and MEPS uses 139–100–2, batch size 16, learning rate 10^{-4}, weight decay 10^{-5}, and 50 epochs. Here BN denotes batch normalization.

I.2 Image and text base predictors

The large-model experiments follow the vision/language setup of the reference benchmark: transformer backbones are fine-tuned from pretrained weights, whereas the ResNet image companion model is trained from scratch as a task-specific classifier. For each named backbone, the preprocessing or tokenizer, backbone size, classifier head, and training grid are fixed across seeds and

weak-observation models, so differences across reported result rows are attributable to the supervision and post-processing data rather than to architecture changes.

The large-model base predictors are as follows. CelebA-ViT predicts the binary facial attribute **`Blond_Hair`** using a pretrained ViT backbone with a binary classification head; the ViT preprocessing is fixed across PN, PU, and UU rows, and PU-trained bases use nnPU. CelebA-ImageResNet predicts the same **`Blond_Hair`** target with a ResNet-50 binary classifier trained as the appendix companion image model. CivilComments-BERT predicts toxicity with a pretrained BERT-based encoder and a binary classification head, using the same tokenizer and classification head across PN and weak-supervision regimes.

CelebA-ViT. The CelebA task is binary classification of the facial attribute **`Blond_Hair`**. The ViT setting uses a pretrained ViT backbone with a binary classification head. Images are resized and normalized according to the ViT preprocessing pipeline, and the model is fine-tuned for 10 epochs with batch size 64, Adam, learning rate 10^{-4}, and weight decay 10^{-2}. The same ViT architecture and image preprocessing are used for PN, PU, and UU rows; weakly trained rows differ only in the weak training objective and observed data view.

CelebA-ImageResNet. The ImageResNet companion experiment uses a ResNet-50 binary classifier for the same **`Blond_Hair`** target. The model is trained for 50 epochs with batch size 64, SGD, momentum 0.9, and learning rate 10^{-3}. Table 29 reports this ResNet experiment in the same format as Tables 28 and 30.

CivilComments-BERT. The CivilComments task is a binary text classification problem: given the text of an online comment, the model predicts whether the comment is toxic or non-toxic. Following the CivilComments setup used in Hansen et al. [21], the identity annotations are not prediction targets; they are used to define evaluation groups for group-wise calibration and multicalibration metrics. Text is tokenized with the corresponding pretrained tokenizer, truncated or padded to maximum length 300, and fine-tuned for 10 epochs with batch size 16, Adam, learning rate 10^{-5}, and weight decay 10^{-2}. The same tokenizer, maximum length, and classification head are reused across PN and weak-supervision regimes.

I.3 Post-processing hyperparameters and WLMC implementation

The post-processing methods use a frozen base predictor. The non-WLMC post-processing optimizers are as follows, while Table 3 gives the WLMC audit/update parameters.

For temperature scaling, the fitted map is a single scalar temperature applied to the logit score. For tabular rows, the objective is clean PN log-loss for PN, corrected weak log-loss for Pconf and UU, and the non-negative PU objective in Eq. (46) for PU, as used in the supplied implementation. We optimize this scalar with L-BFGS-B; tabular weak-temperature rows use at most 200 iterations, and the large-model exported rows use the same fixed post-processing routine used for the corresponding reported result tables.

For Platt scaling, the fitted map is the affine logit transformation $a\,\mathrm{logit}(f(x)) + b$. For tabular rows, the objective is clean PN log-loss for PN, corrected weak log-loss for Pconf and UU, and the non-negative PU objective in Eq. (46) for PU, as used in the supplied implementation. We optimize the two affine parameters with L-BFGS-B and use a maximum of 250 iterations in the tabular exported grid.

Finite-basis WLMC. All real-data WLMC runs use the same finite group–bin witness family as the reported MC metric. The witness family is the Cartesian product of the adopted dataset subgroups and uniform score-bin indicators. Because both $\mathcal{C}$ and $\mathcal{W}$ are finite in all main-text experiments, the audit maximizes over $\mathcal{C} \times \mathcal{W}$ exactly. Consequently, the audit step is implemented by exhaustive enumeration over all subgroup–score-bin cells rather than by an external learning oracle. In each boosting round, WLMC audits all cells under the requested supervision regime (PN, PU, UU, or Pconf), selects the cell with the largest absolute signed violation, and applies an additive correction to the scores in that cell. If the selected cell is A_t and the signed violation is v_t, the update is

$$f_{t+1}(x) = \mathrm{clip}_{[0,1]}\left(f_t(x) + \eta\,\mathrm{sign}(v_t)\mathbf{1}\{x \in A_t\}\right).$$

Table 3: WLMC hyperparameters used in the real-data experiments. All settings use exhaustive auditing over the finite subgroup–score-bin witness family.

Parameter	Tabular	CelebA-ViT	CivilComments-BERT
Number of bins	10	10	10
Step size η	0.05	0.04	0.05
Number of rounds T	50	50	50
Threshold α	0.005	0.005	0.005
Minimum active mass	0.01	0.01	0.01
Positive-confidence $r_{\min}$	10^{-3}	–	–

The procedure stops after T rounds or when the largest absolute violation is below the threshold α. Cells with empirical active mass below the minimum active mass are ignored. For Pconf audits, confidences are lower-clipped at $r_{\min}$ when clipping is enabled.

CelebA-ImageResNet uses the same finite CelebA subgroup family and 10-bin witness construction. Table 29 is a companion to the CelebA-ViT results in Table 28 and now reports test accuracy as well as calibration metrics.

J Reproducibility capsule

All experiments were run on a workstation with two NVIDIA A6000 Ada GPUs and an Intel(R) Xeon(R) w5-3435X CPU running at 3.10 GHz. The released reproduction code records the exact Python environment, package versions, command-line arguments, random seeds, and data-processing scripts used to generate the figures and tables. In this appendix, we therefore specify the experimental protocol, model families, and hardware, while deferring exact package-version strings to the reproducibility package.

The tabular experiments run within approximately 5 hours on CPU/GPU, while the CelebA and CivilComments large-model experiments require approximately 72 hours on GPU. Peak GPU memory usage is approximately 40 GB for both CelebA-ViT and for CivilComments-BERT.

The experiment design follows the clean-label multicalibration benchmark of Hansen et al. [21] for dataset preprocessing, finite subgroup construction, tabular base-model families, and large-model training grids; our changes are the weak-observation views, corrected weak objectives, and weak-label post-processing/evaluation rules described above. The tabular experiments use six base-model families, including MLP. All reported post-processing methods are fit on the correction split, selected on the validation split using the supervision-appropriate validation criterion, and evaluated once on the held-out test split.

K Results on Tabular Datasets

The main text reports diagnostic plots. Here we show the fixed-hyperparameter table format used for detailed inspection. We keep only the representative WLMC implementation setting with step size 0.05 and update cap 50 so that each supervision family appears once, rather than repeating many nearly identical rows for different WLMC budgets.

Coverage of tabular base predictors. Tables 4–27 report the tabular test results for the four datasets used in the main text—ACSIncome, CreditDefault, HMDA, and MEPS—and the six base-predictor families: DecisionTree, LogisticRegression, NaiveBayes, RandomForest, SVM, and MLP. Rows are grouped by $(\mathrm{base}, \mathrm{post})$; the oracle columns use clean test labels for reporting only, the weak columns use the corresponding corrected weak estimates, accuracy is reported against clean test labels, and boldface marks the single best mean within each displayed block.

The tabular scatter diagnostics use the same ten seeds, four datasets, and six base-predictor families. Therefore the WLMC-only and best-of-three post-processing scatters each have $10 \times 4 \times 6 = 240$ points in the $(\mathrm{PN}, \mathrm{PN})$ panel and $10 \times 4 \times 6 \times 3 = 720$ points in each of the $(\mathrm{PN}, \mathrm{weak})$ and $(\mathrm{weak}, \mathrm{weak})$ panels, where the additional factor of 3 is the number of weak-observation types $\{\mathrm{PU}, \mathrm{Pconf}, \mathrm{UU}\}$. The oracle-versus-weak scatter that displays all three post-processing families

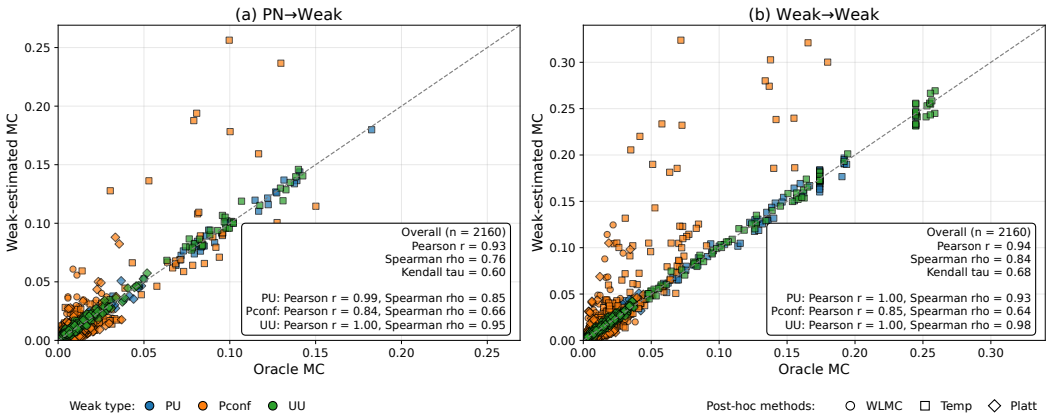


Figure 5: Oracle versus weak-estimated MC on tabular data, with correlation coefficients. The dashed diagonal denotes exact agreement.

has $10 \times 4 \times 6 \times 3 \times 3 = 2160$ points per panel and metric after including the additional factor of 3 for WLMC, temperature scaling, and Platt scaling.

The WLMC-only scatter and the tabular tables report fixed correction methods separately, while Figure 3 uses clean validation labels to select among WLMC, temperature scaling, and Platt scaling using oracle validation MC.

Figures 5 and 6 provide the oracle-versus-weak MC and ECE plots with correlation coefficients.

Tabular result summary. The tabular results show that the proposed WLMC is particularly stable for the MC-oriented quantities: it often reduces subgroup-wise multicalibration error and maxECE without relying on clean labels at post-processing time, and the oracle-versus-weak scatter plots indicate that the weak estimates track the corresponding oracle quantities well enough to guide model selection. Platt scaling is frequently competitive for ECE because it can adjust both slope and intercept of the score distribution, but its effect on the worst subgroup-bin violations is less uniform. Temperature scaling is the most fragile of the three post-processing families: a single global scale can improve average calibration in some easy cases, but it often fails to correct subgroup- or bin-specific errors and can even worsen ECE or MC when the base scores are already misaligned in a non-uniform way. WLMC directly targets the WSL-estimated multicalibration violations, and therefore often continues to improve MC and maxECE even when the one-dimensional score transformations are insufficient.

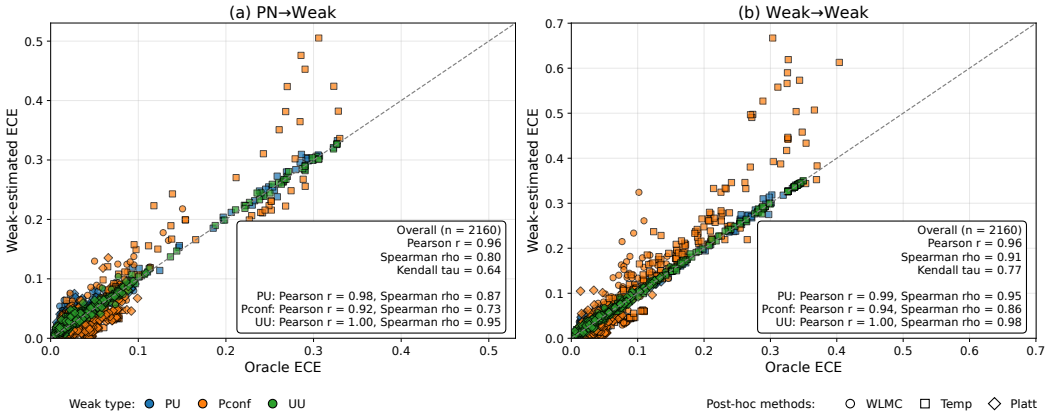


Figure 6: Oracle versus weak-estimated ECE on tabular data, with correlation coefficients. The dashed diagonal denotes exact agreement.

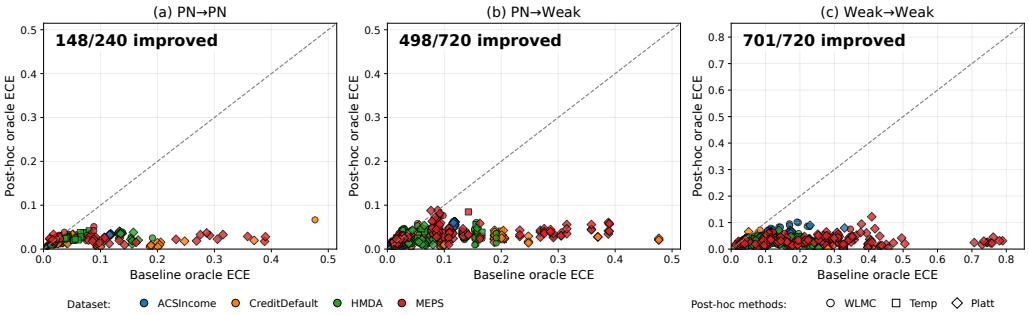


Figure 7: Oracle test ECE before and after correction by a method selected from WLMC, temperature scaling, and Platt scaling using the smallest oracle validation ECE.

Table 4: Tabular test results on ACSIncome with DecisionTree as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**2.48±0.15**	7.42±0.68	0.37±0.02	2.48±0.15	7.43±0.68	0.37±0.02	80.07±0.10
	WLMC	2.48±0.15	7.42±0.68	**0.37±0.02**	**2.48±0.15**	7.43±0.68	**0.37±0.02**	**80.13±0.10**
	Temp	4.35±0.33	9.18±1.12	0.50±0.07	4.35±0.33	9.18±1.12	0.50±0.07	80.13±0.10
	Platt	2.68±0.13	**7.11±1.02**	0.70±0.04	2.68±0.13	**7.10±1.02**	0.70±0.04	80.13±0.11
(PN,PU)	WLMC	**2.46±0.17**	**7.38±0.71**	**0.37±0.02**	**2.72±0.48**	**15.34±2.88**	**0.41±0.10**	**80.14±0.11**
	Temp	4.00±0.33	9.10±1.25	0.47±0.07	4.28±0.74	17.96±4.25	0.50±0.09	80.13±0.10
	Platt	4.07±0.40	8.49±0.90	0.46±0.09	4.28±0.75	17.66±4.52	0.50±0.12	80.13±0.10
(PN,Pconf)	WLMC	**2.39±0.32**	**7.40±0.69**	**0.37±0.02**	**2.93±0.24**	**10.65±6.97**	**0.37±0.15**	**80.13±0.10**
	Temp	3.86±0.68	8.73±0.80	0.43±0.10	3.58±0.64	20.36±13.99	0.55±0.17	80.13±0.10
	Platt	3.94±0.73	8.20±0.95	0.46±0.11	3.68±0.72	19.55±13.32	0.53±0.16	80.11±0.12
(PN,UU)	WLMC	**2.48±0.15**	**7.42±0.68**	**0.37±0.02**	**2.61±0.25**	**11.57±1.85**	**0.43±0.04**	80.13±0.10
	Temp	4.17±0.56	9.26±1.06	0.49±0.09	4.19±0.64	13.99±2.95	0.49±0.07	80.13±0.10
	Platt	4.25±0.61	8.66±0.98	0.48±0.11	4.30±0.71	13.13±3.38	0.49±0.09	**80.13±0.10**
(PU,PU)	ERM	5.60±0.28	12.86±1.09	1.01±0.13	5.77±0.61	17.29±3.01	1.01±0.21	78.41±0.25
	WLMC	**4.17±0.47**	**11.28±0.89**	**0.48±0.12**	**4.29±0.43**	**16.75±4.76**	**0.53±0.11**	78.43±0.25
	Temp	13.45±0.74	14.79±0.95	1.62±0.21	13.47±0.69	18.18±2.23	1.68±0.21	**78.44±0.26**
	Platt	13.56±0.72	15.44±0.86	1.75±0.26	13.62±0.88	18.41±2.27	1.74±0.25	76.11±1.31
(Pconf,Pconf)	ERM	15.48±0.16	22.69±0.54	2.25±0.24	18.18±1.11	108.83±48.36	3.82±0.77	76.34±0.28
	WLMC	7.41±0.52	13.32±0.81	0.61±0.12	8.09±0.86	**23.26±16.37**	**0.67±0.19**	78.20±0.33
	Temp	15.89±0.17	22.28±0.60	1.77±0.16	18.28±0.93	87.43±43.70	3.22±0.72	76.34±0.28
	Platt	**2.47±0.24**	**6.22±0.62**	**0.35±0.05**	**3.15±0.46**	23.87±16.35	0.78±0.22	**80.17±0.19**
(UU,UU)	ERM	3.59±0.16	9.54±0.70	0.58±0.08	3.65±0.29	11.54±1.54	0.57±0.09	**79.39±0.20**
	WLMC	**3.05±0.60**	**9.10±0.68**	**0.42±0.07**	**3.13±0.48**	**11.50±1.36**	**0.42±0.10**	79.39±0.20
	Temp	10.59±1.61	12.10±1.38	1.23±0.27	10.66±1.79	16.31±2.76	1.26±0.29	79.39±0.20
	Platt	10.53±1.55	12.42±1.31	1.15±0.19	10.63±1.71	15.47±2.45	1.17±0.25	79.27±0.30

Table 5: Tabular test results on ACSIncome with LogisticRegression as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**0.93±0.30**	**7.60±1.32**	**0.31±0.04**	**0.93±0.30**	**7.60±1.33**	**0.31±0.04**	**77.02±0.90**
	WLMC	0.93±0.30	7.60±1.32	0.31±0.04	0.93±0.30	7.60±1.33	0.31±0.04	77.02±0.90
	Temp	2.65±0.47	8.81±1.37	0.41±0.06	2.65±0.47	8.81±1.36	0.41±0.06	77.02±0.90
	Platt	2.35±0.11	7.97±1.01	0.52±0.09	2.35±0.11	7.97±1.02	0.52±0.09	77.00±0.92
(PN,PU)	WLMC	**0.92±0.27**	7.64±1.30	**0.30±0.03**	**1.56±0.34**	15.74±3.10	**0.40±0.10**	**77.02±0.90**
	Temp	0.98±0.23	7.79±1.15	0.31±0.03	1.63±0.31	16.24±3.90	0.42±0.11	77.02±0.90
	Platt	0.93±0.23	**7.20±0.68**	0.35±0.05	1.62±0.33	**14.97±3.38**	0.45±0.14	77.00±0.91
(PN,Pconf)	WLMC	1.45±0.49	**7.24±0.88**	0.51±0.19	**1.89±0.34**	**22.17±10.19**	**0.49±0.17**	**77.02±0.90**
	Temp	**1.36±0.45**	7.37±1.18	**0.29±0.04**	1.99±0.26	26.44±13.81	0.50±0.30	77.02±0.90
	Platt	2.28±0.62	7.28±1.57	0.45±0.08	2.18±0.51	24.30±12.50	0.58±0.20	76.90±0.94
(PN,UU)	WLMC	**0.93±0.30**	7.62±1.34	**0.31±0.04**	1.19±0.29	11.83±2.74	**0.33±0.02**	**77.02±0.90**
	Temp	1.00±0.31	7.62±1.29	0.32±0.04	1.23±0.33	11.80±2.44	0.33±0.02	77.02±0.90
	Platt	0.94±0.32	**7.13±0.85**	0.36±0.04	**1.16±0.37**	**11.39±2.12**	0.38±0.05	77.01±0.91
(PU,PU)	ERM	4.52±1.31	14.59±1.08	1.28±0.22	4.75±1.19	20.16±2.48	1.30±0.23	73.63±0.72
	WLMC	**2.28±0.55**	**10.42±1.57**	**0.48±0.08**	**2.58±0.50**	**17.50±4.29**	**0.54±0.07**	**73.78±0.54**
	Temp	4.32±1.45	14.22±1.40	1.17±0.26	4.61±1.44	19.72±2.32	1.21±0.24	73.63±0.72
	Platt	3.82±1.15	12.30±0.85	1.13±0.30	4.14±1.14	18.05±2.38	1.17±0.29	73.69±0.67
(Pconf,Pconf)	ERM	22.68±0.21	35.18±0.36	4.32±0.06	27.39±2.26	235.04±83.69	6.20±0.94	56.20±0.61
	WLMC	6.24±0.78	14.04±1.11	0.69±0.17	7.56±1.01	**34.59±24.13**	**1.19±0.81**	72.91±0.67
	Temp	19.84±1.09	30.81±0.47	7.22±0.38	23.74±2.26	232.29±83.48	9.83±1.82	56.20±0.61
	Platt	**3.04±0.57**	**9.63±0.85**	**0.62±0.07**	**3.53±0.64**	64.39±30.06	1.45±0.50	**74.60±0.41**
(UU,UU)	ERM	3.48±0.97	13.90±0.82	1.10±0.19	3.52±1.04	17.35±2.38	1.10±0.20	74.16±0.49
	WLMC	**1.94±0.42**	**10.46±0.89**	**0.48±0.07**	**2.10±0.45**	**13.90±2.54**	**0.48±0.06**	**74.19±0.39**
	Temp	3.21±1.04	13.14±1.01	0.96±0.20	3.25±1.10	17.04±1.93	0.97±0.21	74.16±0.49
	Platt	2.92±0.84	11.75±0.47	0.91±0.21	2.95±0.90	15.54±1.89	0.91±0.21	74.15±0.46

Table 6: Tabular test results on ACSIncome with NaiveBayes as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	11.73±0.10	16.62±0.18	1.90±0.01	11.73±0.10	16.62±0.18	1.90±0.01	75.34±0.05
	WLMC	6.31±0.21	12.17±0.17	**0.46±0.04**	6.31±0.21	12.17±0.16	**0.46±0.04**	**76.67±0.05**
	Temp	7.76±0.13	15.03±0.18	1.05±0.03	7.76±0.13	15.03±0.18	1.05±0.03	75.34±0.05
	Platt	**3.32±0.07**	**10.02±0.48**	0.63±0.01	**3.32±0.07**	**10.02±0.48**	0.63±0.01	76.52±0.05
(PN,PU)	WLMC	6.37±0.57	12.56±0.53	**0.51±0.17**	**6.25±0.53**	19.64±3.07	**0.58±0.24**	76.55±0.34
	Temp	7.69±0.21	14.66±0.31	0.97±0.03	7.84±0.70	20.14±3.10	1.04±0.08	75.34±0.05
	Platt	**6.14±0.18**	**10.38±0.31**	0.73±0.02	6.35±0.35	**16.24±2.52**	0.76±0.05	**76.65±0.04**
(PN,Pconf)	WLMC	7.18±0.46	11.34±0.71	1.02±0.31	7.61±0.39	**39.25±11.25**	**0.80±0.20**	76.52±0.31
	Temp	7.40±0.19	14.44±0.28	**0.92±0.03**	8.38±0.38	84.16±30.68	1.41±0.28	75.34±0.05
	Platt	**5.48±0.38**	**9.94±0.38**	0.95±0.13	**5.79±0.51**	60.05±21.74	1.01±0.12	**76.80±0.04**
(PN,UU)	WLMC	6.26±0.21	12.19±0.29	**0.45±0.06**	6.23±0.21	15.94±1.91	**0.47±0.06**	76.62±0.10
	Temp	7.72±0.14	14.73±0.21	0.98±0.02	7.64±0.29	18.38±1.95	1.00±0.06	75.34±0.05
	Platt	**6.19±0.13**	**10.23±0.30**	0.72±0.02	**6.13±0.28**	**14.41±1.86**	0.75±0.04	**76.66±0.05**
(PU,PU)	ERM	11.83±0.33	16.60±0.58	1.91±0.02	11.66±0.41	20.98±2.63	1.84±0.06	75.31±0.14
	WLMC	6.53±0.62	12.52±0.57	**0.50±0.19**	**6.41±0.74**	18.82±2.89	**0.59±0.24**	76.48±0.40
	Temp	7.89±0.43	14.64±0.49	0.99±0.06	8.04±0.73	20.00±2.98	1.06±0.08	75.31±0.14
	Platt	**6.36±0.37**	**10.37±0.34**	0.77±0.05	6.55±0.60	**15.74±2.38**	0.79±0.09	**76.66±0.08**
(Pconf,Pconf)	ERM	17.32±0.25	27.74±0.43	2.40±0.07	18.17±0.96	147.43±49.19	2.67±0.28	68.76±0.22
	WLMC	8.50±0.53	13.25±1.08	1.21±0.32	8.47±0.70	**47.96±16.96**	**1.12±0.22**	74.75±0.60
	Temp	13.44±0.19	24.02±0.33	2.44±0.06	14.61±1.38	183.88±65.84	3.00±0.72	68.76±0.22
	Platt	**6.41±0.36**	**9.84±0.40**	**0.82±0.06**	**6.16±0.52**	83.53±29.66	1.25±0.42	**76.33±0.11**
(UU,UU)	ERM	11.72±0.15	16.67±0.33	1.90±0.02	11.68±0.19	18.99±1.70	1.90±0.10	75.29±0.08
	WLMC	**6.16±0.31**	12.08±0.49	**0.42±0.08**	6.12±0.27	15.72±2.02	**0.44±0.08**	76.63±0.13
	Temp	7.73±0.21	14.80±0.27	0.98±0.03	7.64±0.34	18.40±1.93	1.00±0.06	75.29±0.08
	Platt	6.17±0.21	**10.29±0.27**	0.74±0.03	**6.09±0.42**	**14.41±2.03**	0.77±0.05	**76.64±0.06**

Table 7: Tabular test results on ACSIncome with RandomForest as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**0.96±0.10**	**5.39±0.66**	**0.16±0.02**	**0.96±0.10**	5.39±0.66	**0.16±0.02**	**81.70±0.08**
	WLMC	0.96±0.10	5.39±0.66	0.16±0.02	0.96±0.10	**5.39±0.66**	0.16±0.02	81.70±0.08
	Temp	2.64±0.09	6.34±0.67	0.32±0.03	2.64±0.09	6.34±0.67	0.32±0.03	81.70±0.08
	Platt	2.77±0.15	5.88±0.37	0.54±0.03	2.77±0.15	5.88±0.37	0.54±0.03	81.69±0.07
(PN,PU)	WLMC	0.96±0.10	**5.39±0.66**	0.16±0.02	1.56±0.28	**13.88±2.94**	**0.25±0.05**	**81.70±0.08**
	Temp	**0.87±0.18**	5.56±1.00	0.13±0.04	1.53±0.35	14.79±3.04	0.27±0.06	81.70±0.08
	Platt	0.88±0.18	5.53±0.95	**0.13±0.04**	**1.50±0.36**	14.76±3.00	0.26±0.06	81.70±0.10
(PN,Pconf)	WLMC	0.96±0.10	5.39±0.66	0.16±0.02	0.26±0.05	1.74±0.82	0.06±0.04	81.70±0.08
	Temp	**0.96±0.15**	5.40±0.67	0.16±0.03	0.26±0.05	1.73±0.82	0.06±0.04	81.70±0.08
	Platt	0.97±0.13	**5.28±0.68**	**0.16±0.03**	**0.19±0.04**	**1.65±0.77**	**0.05±0.03**	**81.70±0.08**
(PN,UU)	WLMC	0.96±0.10	**5.39±0.66**	0.16±0.02	1.19±0.26	**10.22±2.26**	0.20±0.03	**81.70±0.08**
	Temp	**0.74±0.13**	5.51±0.85	**0.13±0.02**	**1.02±0.30**	10.52±1.74	0.17±0.03	81.70±0.08
	Platt	0.79±0.15	5.50±0.98	0.13±0.03	1.12±0.33	10.59±1.72	**0.16±0.04**	81.69±0.10
(PU,PU)	ERM	20.17±5.56	30.07±5.87	5.95±4.25	20.23±5.57	33.60±6.94	5.99±4.15	73.34±3.89
	WLMC	**9.09±0.67**	17.67±3.11	**0.53±0.08**	**9.19±0.85**	21.61±2.48	**0.62±0.16**	**73.78±3.29**
	Temp	17.04±1.98	25.54±3.23	6.22±1.26	16.92±1.86	29.18±3.86	6.24±1.24	73.34±3.89
	Platt	13.35±3.29	**17.60±2.59**	3.00±1.00	13.24±3.27	**21.33±3.76**	2.99±0.97	68.44±4.07
(Pconf,Pconf)	ERM	17.19±0.17	24.86±0.52	4.01±0.15	19.83±1.60	154.68±60.84	5.65±0.94	77.38±0.16
	WLMC	6.73±0.58	12.17±1.33	0.61±0.14	7.24±0.70	**16.64±6.70**	**0.82±0.29**	79.44±0.34
	Temp	17.02±0.20	23.64±0.30	2.10±0.18	19.39±1.38	104.28±46.53	3.21±0.50	77.38±0.16
	Platt	**4.14±0.16**	**8.59±0.57**	**0.48±0.10**	**4.17±0.43**	28.59±14.99	0.97±0.19	**80.63±0.11**
(UU,UU)	ERM	10.99±0.52	17.06±1.02	2.24±0.18	11.07±0.60	20.74±1.72	2.30±0.24	78.18±0.38
	WLMC	**6.98±0.42**	**13.28±1.12**	**0.49±0.05**	**7.06±0.48**	**17.11±1.62**	**0.51±0.08**	**78.40±0.38**
	Temp	15.48±0.42	21.53±1.38	3.04±0.16	15.37±0.52	25.39±3.75	2.99±0.15	78.18±0.38
	Platt	12.68±0.62	14.83±1.02	1.90±0.09	12.69±0.68	17.17±2.95	1.90±0.12	72.10±0.64

Table 8: Tabular test results on ACSIncome with SVM as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	3.12±0.43	8.16±0.57	0.50±0.05	3.12±0.43	8.15±0.57	0.50±0.05	78.09±0.19
	WLMC	2.74±0.33	7.59±0.44	**0.44±0.05**	2.75±0.33	7.59±0.44	**0.44±0.05**	78.09±0.19
	Temp	10.89±9.50	14.83±9.85	2.82±4.04	10.89±9.50	14.83±9.85	2.82±4.04	66.88±23.67
	Platt	**1.62±0.11**	**7.13±0.43**	0.48±0.03	**1.61±0.11**	**7.13±0.43**	0.48±0.03	**78.17±0.20**
(PN,PU)	WLMC	2.62±0.40	7.66±0.51	0.42±0.08	2.98±0.70	15.35±3.21	0.49±0.09	78.09±0.19
	Temp	1.63±0.50	7.24±0.80	0.31±0.07	1.98±0.36	15.81±3.18	**0.39±0.11**	78.09±0.19
	Platt	**1.25±0.30**	**6.31±0.54**	**0.29±0.04**	**1.74±0.37**	**15.26±3.28**	0.39±0.10	**78.16±0.21**
(PN,Pconf)	WLMC	2.86±0.38	7.35±0.62	0.59±0.22	3.01±0.73	**13.30±8.20**	0.63±0.13	78.09±0.19
	Temp	**1.58±0.56**	7.28±0.56	**0.25±0.07**	2.34±0.91	24.01±11.34	0.62±0.29	78.09±0.19
	Platt	2.05±0.62	**6.14±0.82**	0.31±0.04	**1.82±0.58**	21.20±8.97	**0.58±0.27**	**78.14±0.18**
(PN,UU)	WLMC	2.73±0.41	7.62±0.52	0.44±0.05	2.71±0.40	12.24±2.94	0.45±0.05	78.09±0.19
	Temp	1.69±0.52	7.42±0.73	0.33±0.05	1.82±0.57	11.49±2.17	0.37±0.07	78.09±0.19
	Platt	**1.40±0.28**	**6.38±0.44**	**0.31±0.03**	**1.58±0.33**	**10.77±2.16**	**0.33±0.06**	**78.16±0.20**
(PU,PU)	ERM	3.84±1.21	8.03±1.12	0.58±0.17	4.02±1.43	15.06±3.55	0.64±0.21	78.00±0.39
	WLMC	3.18±0.66	7.30±1.09	0.46±0.05	3.38±0.99	**14.33±2.75**	0.51±0.08	78.09±0.24
	Temp	2.06±0.98	7.53±1.46	0.39±0.20	2.45±0.94	15.12±2.48	0.46±0.22	78.00±0.39
	Platt	**1.54±0.42**	**6.74±0.71**	**0.32±0.11**	**1.98±0.44**	14.99±2.72	**0.39±0.13**	**78.15±0.24**
(Pconf,Pconf)	ERM	23.95±2.03	34.61±3.55	5.35±1.91	27.64±3.42	221.51±74.60	6.32±1.82	55.22±6.60
	WLMC	6.78±0.75	12.34±0.99	0.69±0.21	7.26±1.00	**23.98±8.58**	0.82±0.33	73.06±1.27
	Temp	18.16±5.11	29.49±1.65	6.37±1.69	21.73±4.84	218.11±81.02	8.33±2.93	55.22±6.60
	Platt	**2.51±0.79**	**8.07±1.77**	**0.40±0.11**	**2.16±0.70**	25.06±9.00	**0.55±0.13**	**77.18±0.87**
(UU,UU)	ERM	4.16±0.83	7.95±0.82	0.57±0.08	4.21±0.77	12.80±2.16	0.57±0.11	**78.31±0.11**
	WLMC	3.69±0.51	7.44±0.66	0.50±0.07	3.76±0.46	12.14±1.95	0.50±0.08	78.31±0.11
	Temp	1.77±0.29	6.58±0.89	0.36±0.06	1.91±0.29	10.87±1.85	0.37±0.06	78.31±0.11
	Platt	**1.56±0.22**	**6.44±0.77**	**0.34±0.07**	**1.68±0.25**	**10.80±1.91**	**0.35±0.06**	78.30±0.11

Table 9: Tabular test results on ACSIncome with MLP as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**1.30±0.41**	**6.31±1.09**	0.18±0.06	1.30±0.41	6.31±1.09	0.18±0.06	80.74±0.12
	WLMC	1.30±0.41	6.31±1.09	**0.18±0.06**	**1.30±0.41**	**6.31±1.09**	**0.18±0.06**	80.74±0.12
	Temp	2.86±0.97	6.82±1.04	0.37±0.07	2.86±0.97	6.82±1.04	0.37±0.07	80.74±0.12
	Platt	2.95±0.37	6.95±0.74	0.58±0.09	2.95±0.37	6.95±0.74	0.58±0.09	**80.74±0.13**
(PN,PU)	WLMC	1.30±0.41	**6.31±1.09**	0.18±0.06	**1.55±0.31**	**14.67±3.98**	**0.24±0.05**	80.74±0.12
	Temp	1.10±0.22	6.37±1.12	0.17±0.04	1.61±0.36	15.56±4.11	0.25±0.08	80.74±0.12
	Platt	**0.91±0.09**	6.51±1.04	**0.14±0.04**	1.56±0.41	15.60±4.04	0.25±0.07	**80.75±0.13**
(PN,Pconf)	WLMC	1.34±0.43	6.37±1.02	0.21±0.10	1.43±0.61	5.86±2.78	0.24±0.10	80.74±0.12
	Temp	**1.11±0.21**	6.32±1.13	0.19±0.04	1.25±0.53	5.75±2.50	0.22±0.05	80.74±0.12
	Platt	1.40±0.13	**5.94±0.95**	**0.16±0.03**	**0.77±0.31**	**4.77±2.46**	**0.20±0.05**	**80.74±0.15**
(PN,UU)	WLMC	1.30±0.41	6.31±1.09	0.18±0.06	1.45±0.33	11.40±2.31	0.22±0.07	80.74±0.12
	Temp	1.07±0.18	**6.26±0.92**	0.16±0.05	1.26±0.25	**11.04±1.89**	0.22±0.08	80.74±0.12
	Platt	**0.85±0.17**	6.51±0.92	**0.14±0.03**	**1.22±0.24**	11.30±1.97	**0.20±0.06**	**80.74±0.13**
(PU,PU)	ERM	4.86±3.51	11.08±3.37	0.88±1.05	5.16±3.34	18.27±3.97	0.95±0.98	77.90±1.01
	WLMC	**3.44±1.47**	9.78±1.94	**0.40±0.10**	**3.79±1.45**	16.71±3.73	**0.47±0.11**	**77.94±0.96**
	Temp	4.96±3.44	11.83±4.14	0.70±0.64	5.12±3.46	18.30±4.03	0.74±0.65	77.90±1.01
	Platt	3.88±2.35	**9.48±1.79**	0.45±0.31	4.16±2.41	**16.60±2.83**	0.50±0.32	77.39±2.01
(Pconf,Pconf)	ERM	12.22±1.76	20.03±2.10	2.98±0.76	14.97±2.45	121.09±47.31	4.27±1.04	78.50±0.42
	WLMC	**4.43±0.69**	10.31±1.11	**0.56±0.13**	5.65±1.12	**21.16±10.07**	**1.05±0.60**	**78.83±0.31**
	Temp	11.49±2.27	17.71±2.40	1.59±0.45	13.92±2.87	76.42±29.50	2.40±0.62	78.50±0.42
	Platt	4.74±1.02	**7.99±1.12**	0.65±0.28	**5.02±1.12**	35.22±17.24	1.23±0.37	78.48±0.43
(UU,UU)	ERM	3.10±1.59	8.47±2.06	0.44±0.25	3.20±1.59	12.79±3.22	0.46±0.25	**78.85±0.45**
	WLMC	2.80±0.96	8.12±1.25	0.37±0.10	2.90±0.98	12.37±2.44	0.39±0.09	78.85±0.45
	Temp	2.26±0.87	8.41±2.00	0.33±0.15	2.36±0.92	12.90±2.81	0.36±0.18	78.85±0.45
	Platt	**1.68±0.71**	**7.62±1.22**	**0.28±0.09**	**1.84±0.84**	**11.99±2.48**	**0.29±0.11**	78.77±0.50

Table 10: Tabular test results on CreditDefault with DecisionTree as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	5.42±0.48	16.38±3.61	1.13±0.17	5.43±0.48	16.38±3.61	1.13±0.17	**80.55±0.32**
	WLMC	3.47±0.35	14.98±4.39	**0.52±0.10**	3.47±0.34	14.97±4.39	**0.52±0.10**	80.54±0.24
	Temp	12.20±0.61	17.03±2.18	3.34±0.70	12.20±0.61	17.03±2.18	3.34±0.70	80.54±0.24
	Platt	**2.88±0.44**	**8.70±2.12**	0.97±0.21	**2.88±0.44**	**8.70±2.11**	0.97±0.21	80.43±0.24
(PN,PU)	WLMC	**3.54±0.92**	14.81±4.62	**0.71±0.36**	**4.01±1.05**	24.96±9.06	**0.83±0.34**	**80.54±0.24**
	Temp	22.46±2.56	24.82±2.02	8.18±1.40	22.68±2.54	32.37±7.30	8.14±1.37	80.54±0.24
	Platt	5.72±2.53	**9.92±2.83**	1.98±0.78	5.48±2.64	**14.31±5.41**	1.92±0.78	78.58±0.42
(PN,Pconf)	WLMC	**4.50±0.80**	16.42±5.03	**1.08±0.57**	4.29±0.57	7.59±0.93	**0.62±0.26**	**80.54±0.24**
	Temp	24.55±1.20	26.80±1.03	8.74±2.90	21.67±1.03	28.42±1.54	7.54±1.84	80.54±0.24
	Platt	4.69±1.13	**7.46±1.32**	2.12±0.65	**3.22±0.70**	**6.36±1.22**	1.08±0.30	78.74±0.36
(PN,UU)	WLMC	**3.64±0.33**	14.11±3.35	**0.61±0.12**	**4.11±0.50**	18.43±5.65	**0.79±0.12**	**80.54±0.24**
	Temp	22.11±1.80	24.38±1.44	8.20±0.97	22.15±1.70	28.77±4.50	8.28±0.83	80.54±0.24
	Platt	5.60±2.22	**9.11±2.81**	1.82±0.77	5.47±2.08	**11.37±3.28**	1.82±0.75	78.74±0.36
(PU,PU)	ERM	8.75±1.26	17.29±3.92	2.05±0.26	9.22±1.45	25.99±6.70	2.21±0.43	79.28±0.57
	WLMC	**4.82±0.89**	13.96±4.04	**0.84±0.36**	**5.11±1.07**	25.06±7.03	**0.99±0.29**	**79.31±0.58**
	Temp	25.10±1.16	28.22±1.86	10.43±2.61	25.52±1.35	37.21±9.06	10.40±2.75	79.29±0.57
	Platt	6.56±1.20	**10.78±1.63**	2.31±0.27	6.84±1.36	**18.55±5.85**	2.46±0.45	79.23±0.62
(Pconf,Pconf)	ERM	4.39±0.31	**10.15±1.69**	1.25±0.12	3.88±0.22	5.04±0.47	0.70±0.07	**81.53±0.20**
	WLMC	3.17±0.32	10.50±2.65	1.01±0.32	2.89±0.23	4.22±0.44	0.43±0.09	81.53±0.20
	Temp	3.22±0.33	13.12±1.96	0.94±0.18	3.04±0.18	3.88±0.34	0.42±0.03	81.53±0.20
	Platt	**2.34±0.33**	11.35±2.36	**0.89±0.14**	**0.73±0.17**	**2.22±0.49**	**0.23±0.05**	81.44±0.25
(UU,UU)	ERM	4.84±0.91	14.76±3.34	1.28±0.26	5.43±0.95	19.57±5.39	1.34±0.26	**80.41±0.67**
	WLMC	**3.35±0.64**	13.82±4.09	**0.65±0.16**	**3.66±0.68**	19.29±5.39	**0.75±0.28**	80.41±0.67
	Temp	17.98±2.85	22.21±2.78	5.56±1.20	18.01±2.92	26.04±5.70	5.70±1.40	80.41±0.67
	Platt	8.31±1.07	**11.85±1.82**	2.59±0.30	8.15±1.23	**13.84±2.81**	2.52±0.24	79.04±0.39

Table 11: Tabular test results on CreditDefault with LogisticRegression as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**1.23±0.22**	11.52±3.41	**0.23±0.04**	**1.23±0.22**	11.52±3.41	**0.23±0.04**	**81.84±0.09**
	WLMC	1.26±0.23	11.44±3.50	0.24±0.05	1.27±0.23	11.44±3.49	0.24±0.05	81.84±0.09
	Temp	3.38±0.37	9.60±2.69	1.05±0.16	3.38±0.37	9.60±2.69	1.05±0.16	81.84±0.09
	Platt	2.26±0.24	**7.64±0.81**	0.48±0.16	2.26±0.24	**7.64±0.81**	0.48±0.16	81.81±0.11
(PN,PU)	WLMC	**1.33±0.29**	11.88±3.00	0.30±0.14	2.61±0.91	24.01±7.35	0.63±0.12	**81.86±0.11**
	Temp	1.34±0.29	**10.84±3.08**	0.30±0.11	2.44±0.69	**23.43±7.98**	**0.53±0.11**	81.84±0.09
	Platt	1.54±0.42	11.05±3.92	**0.29±0.10**	**2.36±0.84**	24.99±10.19	0.54±0.09	81.82±0.17
(PN,Pconf)	WLMC	**1.30±0.29**	11.44±3.49	**0.27±0.14**	2.02±0.24	3.80±0.36	0.30±0.07	**81.84±0.09**
	Temp	2.42±0.35	9.30±2.14	0.67±0.13	2.17±0.16	4.07±0.40	0.26±0.05	81.84±0.09
	Platt	2.74±0.39	**7.86±1.60**	0.63±0.11	**0.95±0.15**	**2.85±0.53**	**0.19±0.03**	81.60±0.17
(PN,UU)	WLMC	1.40±0.30	11.61±3.66	0.32±0.16	1.82±0.77	21.03±4.20	0.43±0.11	**81.84±0.09**
	Temp	**1.36±0.28**	10.80±2.11	**0.28±0.07**	1.69±0.62	20.48±4.40	0.37±0.06	81.84±0.09
	Platt	1.44±0.35	**10.38±2.51**	0.28±0.07	**1.63±0.61**	**19.95±5.29**	**0.34±0.08**	81.83±0.10
(PU,PU)	ERM	6.78±0.45	**9.97±1.30**	2.30±0.15	7.05±0.64	19.30±5.37	2.26±0.30	78.12±0.00
	WLMC	2.76±0.72	10.62±3.46	0.64±0.19	3.27±0.77	18.82±5.38	0.86±0.29	79.17±0.49
	Temp	7.51±0.43	10.41±0.69	1.78±0.17	7.72±1.31	**17.81±4.50**	1.92±0.40	78.12±0.00
	Platt	**1.66±0.28**	12.90±2.42	**0.39±0.09**	**2.68±0.95**	25.80±5.33	**0.65±0.09**	**80.62±0.24**
(Pconf,Pconf)	ERM	11.46±0.32	13.18±0.77	5.51±0.20	9.31±0.41	13.23±1.08	4.01±0.23	78.13±0.03
	WLMC	5.03±0.75	9.66±0.94	**0.80±0.43**	3.51±0.53	10.91±0.90	0.54±0.16	78.75±0.34
	Temp	6.68±0.48	**9.39±0.89**	1.92±0.09	5.63±0.33	8.10±0.42	1.10±0.06	78.13±0.03
	Platt	**2.75±0.34**	14.00±1.27	0.85±0.15	**1.46±0.15**	**5.18±0.40**	**0.32±0.04**	**80.44±0.07**
(UU,UU)	ERM	6.66±0.19	9.76±1.12	2.32±0.12	6.41±0.33	13.00±2.08	2.23±0.24	78.12±0.00
	WLMC	2.98±0.51	**9.41±2.99**	0.60±0.22	2.98±0.47	14.80±3.79	0.66±0.11	79.10±0.33
	Temp	7.63±0.37	10.31±0.83	1.80±0.17	7.29±0.57	**12.80±1.59**	1.75±0.27	78.12±0.00
	Platt	**1.64±0.33**	12.90±1.63	**0.34±0.04**	**2.10±0.56**	20.37±4.42	**0.41±0.07**	**80.58±0.17**

Table 12: Tabular test results on CreditDefault with NaiveBayes as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	24.69±9.78	33.00±11.66	10.83±3.93	24.69±9.78	32.99±11.66	10.83±3.93	73.80±11.27
	WLMC	6.31±1.29	13.99±1.25	1.06±0.41	6.31±1.29	13.99±1.24	1.06±0.41	78.72±2.02
	Temp	23.64±8.93	31.81±10.64	9.24±2.14	23.65±8.93	31.81±10.64	9.24±2.14	73.80±11.27
	Platt	**1.98±2.66**	**7.21±2.08**	**0.90±0.68**	**1.97±2.66**	**7.21±2.08**	**0.90±0.68**	**79.43±1.04**
(PN,PU)	WLMC	6.51±1.05	15.59±1.96	1.26±0.50	6.62±1.05	20.75±6.81	1.55±0.56	78.60±2.24
	Temp	11.33±12.02	16.00±11.01	3.33±3.65	11.71±11.80	25.18±10.55	3.40±3.48	73.80±11.27
	Platt	**2.53±0.48**	**8.10±1.50**	**0.78±0.30**	**2.85±0.61**	**18.01±5.58**	**1.00±0.34**	**79.97±0.91**
(PN,Pconf)	WLMC	5.92±1.64	13.30±2.76	**1.25±0.86**	6.02±1.73	11.51±2.32	0.94±0.49	78.55±2.21
	Temp	12.59±11.09	16.08±10.56	4.52±3.72	11.04±10.39	15.44±11.73	3.28±3.35	73.80±11.27
	Platt	**3.26±0.79**	**7.50±2.24**	1.39±0.56	**2.10±0.70**	**5.56±2.08**	**0.62±0.19**	**79.62±1.14**
(PN,UU)	WLMC	6.67±1.49	14.04±1.62	1.06±0.27	6.79±1.61	17.56±3.16	1.12±0.29	78.60±2.24
	Temp	11.50±11.88	16.02±10.87	3.38±3.57	11.53±11.82	19.84±10.51	3.51±3.81	73.80±11.27
	Platt	**2.51±0.64**	**8.08±1.72**	**0.83±0.38**	**2.37±0.66**	**11.25±2.24**	**0.83±0.39**	**79.96±0.92**
(PU,PU)	ERM	28.12±0.00	31.41±0.00	17.40±0.00	28.12±0.00	34.49±6.79	17.14±0.80	21.88±0.00
	WLMC	5.08±1.06	12.68±2.04	1.32±0.57	5.09±1.09	16.74±3.98	1.30±0.47	**78.12±0.00**
	Temp	28.12±0.00	31.41±0.00	17.40±0.00	28.12±0.00	34.49±6.79	17.14±0.80	78.12±0.00
	Platt	**0.45±0.39**	**5.87±0.48**	**0.88±0.22**	**0.45±0.39**	**9.05±4.51**	**1.14±0.32**	78.12±0.00
(Pconf,Pconf)	ERM	20.03±0.25	27.03±0.53	9.96±0.22	20.07±0.17	25.79±0.48	9.96±0.22	**79.73±0.21**
	WLMC	5.49±1.09	11.64±1.38	**1.53±0.82**	4.48±0.82	10.74±0.60	**0.80±0.57**	79.73±0.21
	Temp	4.95±0.85	9.17±1.53	2.34±0.39	3.55±0.67	**7.67±1.83**	1.18±0.24	79.73±0.21
	Platt	**4.20±0.44**	**8.44±1.50**	1.99±0.19	**2.87±0.56**	7.87±2.00	0.87±0.13	79.41±0.28
(UU,UU)	ERM	28.12±0.00	31.41±0.00	17.40±0.00	28.12±0.00	33.67±3.99	17.51±0.39	21.88±0.00
	WLMC	4.77±1.33	12.68±2.13	1.17±0.65	4.67±1.27	16.70±4.54	1.22±0.60	**78.12±0.00**
	Temp	28.12±0.00	31.41±0.00	17.40±0.00	28.12±0.00	33.67±3.99	17.51±0.39	78.12±0.00
	Platt	**0.45±0.39**	**5.87±0.48**	**0.88±0.22**	**0.45±0.39**	**6.89±1.71**	**1.07±0.29**	78.12±0.00

Table 13: Tabular test results on CreditDefault with RandomForest as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	1.76±0.27	11.91±3.74	0.43±0.10	1.76±0.27	11.92±3.74	0.43±0.10	81.81±0.18
	WLMC	**1.55±0.32**	11.74±3.24	**0.37±0.09**	**1.55±0.32**	11.74±3.24	**0.37±0.09**	81.81±0.18
	Temp	6.85±9.57	10.44±9.26	3.03±5.81	6.85±9.57	10.44±9.27	3.03±5.81	75.45±20.11
	Platt	1.96±0.21	**8.80±2.30**	0.46±0.14	1.95±0.21	**8.80±2.29**	0.46±0.14	**81.90±0.16**
(PN,PU)	WLMC	1.50±0.51	**11.78±3.82**	0.46±0.28	**2.73±0.78**	**23.51±5.31**	0.73±0.20	81.81±0.22
	Temp	1.52±0.46	13.11±3.52	**0.30±0.06**	2.74±0.71	26.59±8.62	0.71±0.28	81.81±0.18
	Platt	**1.44±0.54**	13.87±3.90	0.31±0.09	3.02±0.76	28.97±8.48	**0.70±0.23**	**81.88±0.10**
(PN,Pconf)	WLMC	**1.76±0.27**	11.91±3.74	**0.43±0.10**	0.23±0.05	**0.83±0.12**	0.05±0.01	**81.81±0.18**
	Temp	1.80±0.24	11.88±3.33	0.43±0.09	0.22±0.04	0.84±0.13	0.05±0.01	81.81±0.18
	Platt	1.84±0.27	**11.88±3.28**	0.44±0.10	**0.19±0.06**	0.85±0.16	**0.05±0.01**	81.80±0.17
(PN,UU)	WLMC	1.53±0.35	**12.40±4.01**	0.36±0.09	2.06±0.50	**19.65±6.47**	0.53±0.13	81.81±0.18
	Temp	1.35±0.20	12.86±2.64	0.29±0.06	**1.80±0.30**	20.49±4.02	0.48±0.15	81.81±0.18
	Platt	**1.24±0.35**	13.91±3.19	**0.29±0.06**	2.03±0.55	19.67±4.34	**0.48±0.14**	**81.89±0.16**
(PU,PU)	ERM	7.58±3.95	20.26±5.49	2.65±2.72	8.00±4.04	30.08±7.16	2.57±2.44	79.00±2.56
	WLMC	**4.08±0.74**	17.74±3.61	**0.59±0.14**	4.66±1.23	28.84±7.17	**0.92±0.24**	**79.67±1.56**
	Temp	18.92±10.00	23.83±8.34	8.11±6.78	19.29±9.97	35.80±8.46	8.13±6.64	79.00±2.56
	Platt	4.60±2.61	**10.47±3.28**	1.38±0.70	**4.65±2.84**	**22.70±9.76**	1.45±0.89	78.77±1.79
(Pconf,Pconf)	ERM	8.04±0.26	15.99±2.23	2.36±0.20	6.36±0.40	8.37±0.82	1.89±0.19	**81.33±0.21**
	WLMC	4.11±0.79	17.89±2.73	**0.84±0.19**	3.07±0.42	5.41±0.78	0.58±0.15	81.33±0.21
	Temp	4.15±0.36	12.69±1.81	1.17±0.10	3.27±0.30	4.79±0.77	0.53±0.08	81.33±0.21
	Platt	**3.59±0.39**	**11.94±1.96**	0.90±0.10	**1.94±0.23**	**3.89±0.41**	**0.39±0.09**	80.96±0.23
(UU,UU)	ERM	8.16±0.55	17.93±4.04	1.63±0.27	8.48±0.64	25.92±7.62	1.83±0.35	79.34±0.45
	WLMC	**4.90±1.19**	15.74±3.24	**0.71±0.16**	**5.29±1.32**	22.23±4.85	**0.83±0.18**	**79.52±0.38**
	Temp	21.67±2.34	24.56±2.15	6.98±1.55	21.91±2.44	31.12±7.75	7.09±1.56	79.34±0.45
	Platt	8.16±0.73	**10.51±1.07**	2.25±0.32	8.21±0.96	**14.10±3.41**	2.26±0.40	79.05±0.28

Table 14: Tabular test results on CreditDefault with SVM as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	9.64±0.25	11.92±0.49	5.81±0.09	9.64±0.25	11.92±0.49	5.81±0.09	81.88±0.05
	WLMC	2.85±0.77	7.21±0.90	0.86±0.43	2.85±0.77	7.21±0.90	0.86±0.43	81.88±0.05
	Temp	15.34±0.08	17.14±0.19	9.37±0.08	15.33±0.08	17.14±0.19	9.37±0.08	81.88±0.05
	Platt	**0.76±0.31**	**5.33±0.70**	**0.61±0.23**	**0.75±0.31**	**5.34±0.70**	**0.61±0.23**	**81.88±0.05**
(PN,PU)	WLMC	3.92±1.16	8.51±1.24	1.56±0.71	4.30±1.25	17.91±7.51	1.64±0.70	81.88±0.05
	Temp	3.64±1.70	7.93±1.63	1.47±0.71	3.87±2.22	19.58±8.73	1.59±0.77	81.88±0.05
	Platt	**1.45±0.66**	**5.82±1.33**	**0.68±0.24**	**1.56±0.53**	**16.90±10.11**	**0.82±0.15**	**81.88±0.05**
(PN,Pconf)	WLMC	**3.10±0.83**	7.79±0.91	**1.07±0.75**	3.42±0.56	8.32±1.40	0.74±0.44	81.88±0.05
	Temp	5.20±0.52	8.40±0.59	2.59±0.60	4.94±0.37	7.81±0.77	1.56±0.33	81.88±0.05
	Platt	3.23±1.54	**5.81±1.49**	1.58±0.55	**1.28±1.19**	**4.59±1.31**	**0.61±0.31**	**81.88±0.05**
(PN,UU)	WLMC	3.16±1.11	7.72±1.23	1.17±0.69	3.42±1.20	11.30±3.89	1.23±0.70	81.88±0.05
	Temp	4.01±1.56	8.12±2.29	1.65±0.58	4.13±1.70	11.85±5.98	1.70±0.63	81.88±0.05
	Platt	**0.99±0.39**	**5.78±1.31**	**0.66±0.26**	**1.17±0.45**	**9.46±3.66**	**0.74±0.32**	**81.88±0.05**
(PU,PU)	ERM	9.86±0.36	13.71±2.31	5.67±0.21	10.25±0.77	21.78±8.91	5.65±0.38	**81.91±0.05**
	WLMC	3.43±1.10	9.56±2.22	**1.01±0.67**	3.83±1.15	21.17±7.89	1.21±0.63	81.91±0.05
	Temp	4.23±1.29	9.28±1.52	1.57±0.56	4.45±1.26	22.97±8.38	1.80±0.46	81.91±0.05
	Platt	**2.99±2.12**	**7.10±1.40**	1.06±0.72	**3.18±2.14**	**18.29±8.97**	**1.19±0.51**	81.20±1.49
(Pconf,Pconf)	ERM	10.29±0.35	13.62±1.42	5.94±0.15	9.41±0.57	13.88±1.22	4.46±0.34	**81.93±0.03**
	WLMC	3.94±0.97	9.84±1.73	**1.37±0.74**	3.69±0.73	10.39±1.19	**0.75±0.54**	81.93±0.03
	Temp	4.96±0.55	8.71±2.29	1.84±0.34	4.78±0.64	8.78±1.17	1.21±0.26	81.93±0.03
	Platt	**3.90±0.91**	**6.22±0.90**	1.78±0.32	**2.27±0.83**	**5.36±1.26**	0.75±0.19	81.92±0.02
(UU,UU)	ERM	9.78±0.31	12.37±0.80	5.87±0.09	9.96±0.32	15.03±3.14	5.91±0.27	**81.91±0.04**
	WLMC	3.43±1.28	8.42±1.31	1.28±0.71	3.80±1.32	12.17±3.67	1.40±0.69	81.91±0.04
	Temp	4.13±1.35	8.75±2.83	1.77±0.58	4.14±1.43	12.36±4.54	1.81±0.60	81.91±0.04
	Platt	**1.16±0.61**	**5.62±1.00**	**0.75±0.27**	**1.34±0.52**	**9.58±3.17**	**0.82±0.32**	81.89±0.04

Table 15: Tabular test results on CreditDefault with MLP as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	2.86±1.77	10.61±1.86	0.71±0.62	2.86±1.77	10.61±1.86	0.71±0.62	81.91±0.15
	WLMC	2.07±0.48	10.58±1.82	**0.39±0.12**	2.07±0.48	10.58±1.82	**0.39±0.12**	81.91±0.15
	Temp	5.35±3.71	10.80±3.65	1.54±1.24	5.35±3.72	10.80±3.65	1.54±1.24	81.91±0.15
	Platt	**1.84±0.59**	**9.09±2.66**	0.44±0.18	**1.84±0.59**	**9.09±2.66**	0.44±0.18	**82.04±0.09**
(PN,PU)	WLMC	2.07±0.53	**10.62±2.21**	0.47±0.20	2.86±1.11	**22.77±7.74**	**0.60±0.13**	81.91±0.15
	Temp	1.90±0.61	11.33±2.98	0.39±0.12	2.87±1.08	26.59±6.58	0.61±0.14	81.91±0.15
	Platt	**1.87±0.67**	11.46±3.54	**0.37±0.13**	**2.64±0.95**	27.30±6.77	0.64±0.17	**82.00±0.18**
(PN,Pconf)	WLMC	**2.09±0.57**	9.95±1.56	**0.52±0.22**	2.07±0.68	4.15±1.37	0.36±0.10	**81.91±0.15**
	Temp	2.44±0.54	10.15±3.40	0.67±0.27	2.11±0.74	4.29±1.14	0.34±0.13	81.91±0.15
	Platt	2.62±0.60	**7.85±2.82**	0.60±0.20	**1.20±0.38**	**3.32±0.65**	**0.26±0.06**	81.78±0.19
(PN,UU)	WLMC	2.24±0.32	10.54±1.82	0.50±0.12	2.69±0.40	18.52±6.38	0.60±0.15	81.91±0.15
	Temp	1.84±0.49	10.47±2.93	0.40±0.10	2.22±0.62	**18.52±6.76**	0.48±0.13	81.91±0.15
	Platt	**1.65±0.63**	**10.22±3.85**	**0.37±0.09**	**2.01±0.64**	18.54±5.37	**0.45±0.12**	**82.01±0.10**
(PU,PU)	ERM	3.45±1.64	11.58±3.09	0.91±0.50	4.12±1.36	25.62±7.93	1.09±0.38	**81.32±0.20**
	WLMC	2.18±0.82	12.79±3.65	0.59±0.44	2.92±0.85	25.95±10.09	0.77±0.41	81.32±0.20
	Temp	2.44±1.12	10.97±3.07	0.52±0.19	3.28±1.15	23.10±8.13	0.73±0.23	81.32±0.20
	Platt	**2.12±0.70**	**10.25±1.99**	**0.51±0.19**	**2.81±0.97**	**20.73±7.37**	**0.71±0.21**	81.29±0.32
(Pconf,Pconf)	ERM	4.76±2.78	11.13±2.55	1.55±1.17	4.35±2.60	5.94±3.04	1.08±0.93	**81.68±0.19**
	WLMC	3.09±0.89	**10.67±2.33**	0.83±0.31	2.99±0.94	4.57±1.39	0.45±0.09	81.68±0.19
	Temp	3.06±0.99	12.92±3.43	0.89±0.29	2.96±1.05	4.15±1.24	0.48±0.19	81.68±0.19
	Platt	**2.63±0.45**	13.31±3.83	**0.82±0.21**	**1.20±0.45**	**2.59±0.56**	**0.31±0.09**	81.57±0.23
(UU,UU)	ERM	3.76±1.89	11.89±2.95	0.90±0.40	3.91±2.21	18.58±3.05	0.99±0.45	81.54±0.28
	WLMC	2.41±0.75	12.21±3.05	0.48±0.19	2.66±0.94	17.85±2.68	**0.53±0.18**	81.54±0.28
	Temp	2.58±1.02	11.81±3.79	0.54±0.19	2.88±1.21	**17.56±3.44**	0.67±0.24	81.54±0.28
	Platt	**2.24±0.57**	**11.74±4.43**	**0.47±0.15**	**2.43±0.59**	19.78±5.17	0.54±0.14	**81.56±0.22**

Table 16: Tabular test results on HMDA with DecisionTree as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	2.73±0.17	7.04±1.35	0.44±0.05	2.73±0.17	7.04±1.35	0.44±0.05	86.02±0.25
	WLMC	**2.67±0.19**	**6.99±1.24**	0.41±0.08	**2.67±0.19**	**6.99±1.24**	0.41±0.08	**86.03±0.24**
	Temp	2.81±0.26	8.19±0.80	**0.40±0.08**	2.81±0.26	8.19±0.80	**0.40±0.08**	86.03±0.24
	Platt	4.67±0.28	7.97±1.51	0.98±0.09	4.67±0.29	7.97±1.51	0.98±0.09	85.75±0.29
(PN,PU)	WLMC	**2.63±0.22**	**6.94±1.33**	**0.46±0.12**	**2.74±0.47**	**15.45±2.67**	**0.69±0.31**	**86.03±0.24**
	Temp	3.64±1.42	8.79±1.68	0.65±0.29	3.75±1.34	17.97±2.79	0.75±0.20	86.03±0.24
	Platt	3.56±1.36	8.56±1.45	0.65±0.29	3.67±1.24	17.46±3.15	0.76±0.21	86.00±0.21
(PN,Pconf)	WLMC	**4.38±0.17**	**8.29±1.55**	1.70±0.04	3.92±0.09	6.45±0.42	**0.64±0.05**	**86.03±0.24**
	Temp	7.56±0.19	11.56±1.40	**1.38±0.08**	3.46±0.20	**4.74±0.40**	0.98±0.18	86.03±0.24
	Platt	7.05±0.39	11.90±0.65	1.60±0.08	**3.24±0.24**	4.91±0.30	0.79±0.07	85.31±0.46
(PN,UU)	WLMC	**2.64±0.16**	**6.95±1.27**	**0.40±0.07**	**2.70±0.34**	**9.86±2.74**	**0.43±0.09**	**86.03±0.24**
	Temp	4.24±0.40	9.60±0.66	0.76±0.15	4.19±0.46	12.08±2.81	0.81±0.14	86.03±0.24
	Platt	4.11±0.45	8.98±0.44	0.80±0.13	4.05±0.47	12.19±2.96	0.80±0.12	86.02±0.22
(PU,PU)	ERM	4.36±0.52	9.23±2.72	0.77±0.14	4.45±1.17	**17.77±5.58**	0.96±0.37	84.50±0.46
	WLMC	**3.92±0.65**	**8.87±2.80**	**0.72±0.12**	**4.03±0.72**	17.84±5.52	**0.86±0.34**	**84.53±0.46**
	Temp	5.39±2.12	13.91±3.41	1.59±0.98	5.34±2.37	19.58±4.38	1.69±1.13	84.51±0.47
	Platt	5.28±2.28	14.41±3.76	1.60±0.82	5.39±2.59	20.01±5.34	1.68±0.97	84.19±0.57
(Pconf,Pconf)	ERM	8.61±0.22	18.53±0.84	1.24±0.07	5.34±0.11	7.60±0.31	0.70±0.03	82.01±0.31
	WLMC	7.45±0.23	16.86±0.68	**0.96±0.07**	4.34±0.19	6.71±0.18	0.47±0.03	82.99±0.28
	Temp	9.11±0.26	19.03±0.90	1.37±0.07	4.56±0.09	6.70±0.31	0.74±0.03	82.01±0.31
	Platt	**4.43±0.19**	**12.80±0.95**	1.00±0.07	**0.36±0.07**	**1.87±0.31**	**0.10±0.02**	**86.02±0.14**
(UU,UU)	ERM	2.67±0.30	8.51±0.79	0.62±0.18	2.76±0.42	11.42±2.33	0.67±0.25	**85.60±0.33**
	WLMC	**2.36±0.53**	**8.28±0.85**	**0.42±0.08**	**2.48±0.59**	**11.39±2.64**	**0.46±0.12**	85.60±0.33
	Temp	5.52±0.86	10.41±1.82	0.92±0.19	5.50±0.95	13.31±3.25	0.97±0.19	85.60±0.33
	Platt	5.56±0.93	11.04±1.86	0.96±0.22	5.56±1.03	14.09±3.39	1.01±0.26	85.25±0.68

Table 17: Tabular test results on HMDA with LogisticRegression as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	1.50±0.19	**10.06±0.21**	**0.26±0.03**	1.49±0.19	**10.06±0.22**	**0.26±0.03**	**82.68±0.11**
	WLMC	1.50±0.19	10.06±0.21	0.26±0.03	1.50±0.19	10.06±0.22	0.26±0.03	82.68±0.11
	Temp	**1.10±0.08**	10.17±0.18	0.31±0.02	**1.10±0.08**	10.17±0.18	0.31±0.02	82.68±0.11
	Platt	5.13±0.17	13.40±0.54	0.79±0.06	5.13±0.17	13.40±0.55	0.79±0.06	82.61±0.13
(PN,PU)	WLMC	**1.48±0.20**	**10.06±0.26**	**0.26±0.03**	**2.30±0.52**	**17.22±2.15**	**0.72±0.19**	82.68±0.11
	Temp	2.00±0.43	10.23±0.29	0.40±0.16	2.89±0.67	17.55±2.81	0.89±0.39	82.68±0.11
	Platt	2.44±0.92	10.40±0.70	0.42±0.14	3.22±0.89	17.80±2.51	0.76±0.30	**82.68±0.13**
(PN,Pconf)	WLMC	**1.49±0.18**	**10.06±0.22**	**0.26±0.03**	**3.98±0.17**	6.38±0.29	**1.03±0.06**	**82.68±0.11**
	Temp	5.98±0.17	12.24±0.24	1.23±0.06	4.75±0.17	**6.00±0.21**	1.05±0.05	82.68±0.11
	Platt	7.44±0.22	14.81±0.28	1.44±0.02	4.76±0.15	6.39±0.28	1.19±0.02	80.79±0.30
(PN,UU)	WLMC	**1.50±0.19**	10.06±0.21	**0.26±0.03**	**1.50±0.26**	**11.52±2.09**	**0.32±0.05**	82.68±0.11
	Temp	1.54±0.40	10.08±0.19	0.28±0.04	1.57±0.46	11.69±2.37	0.34±0.05	82.68±0.11
	Platt	1.57±0.37	**10.05±0.30**	0.28±0.04	1.57±0.43	11.62±2.02	0.35±0.08	**82.68±0.13**
(PU,PU)	ERM	3.00±1.61	14.96±1.84	1.88±0.39	3.19±1.46	22.16±5.94	1.91±0.51	75.41±0.04
	WLMC	**1.76±0.74**	**14.23±1.08**	**0.71±0.15**	**2.20±0.69**	23.48±5.66	**0.99±0.35**	**75.43±0.09**
	Temp	2.93±1.53	14.66±1.82	1.85±0.36	3.13±1.40	22.03±5.67	1.87±0.51	75.41±0.04
	Platt	2.85±1.60	15.02±1.32	1.86±0.43	3.01±1.53	**21.79±5.08**	1.89±0.61	75.40±0.06
(Pconf,Pconf)	ERM	10.51±0.15	26.46±0.98	4.32±0.58	7.78±0.20	13.52±0.41	3.40±0.59	**75.44±0.00**
	WLMC	3.61±1.21	18.44±1.61	**1.22±0.41**	1.65±0.53	6.57±0.80	**0.81±0.54**	75.44±0.00
	Temp	2.79±0.81	18.25±0.78	1.49±0.10	**1.11±1.02**	6.16±0.32	1.65±0.41	75.44±0.00
	Platt	**2.44±0.44**	**17.76±1.53**	1.29±0.27	1.37±1.46	**5.94±0.35**	1.70±0.35	75.44±0.00
(UU,UU)	ERM	2.53±1.03	14.14±1.25	1.76±0.29	2.65±0.95	13.54±1.94	1.81±0.32	75.40±0.04
	WLMC	**1.63±0.58**	14.36±0.97	**0.56±0.07**	**1.64±0.64**	**13.51±1.91**	**0.58±0.12**	**75.40±0.04**
	Temp	2.55±0.97	14.41±1.60	1.79±0.38	2.63±0.90	14.46±2.27	1.84±0.38	75.40±0.04
	Platt	2.55±1.03	**14.05±1.46**	1.80±0.36	2.66±1.00	13.85±2.27	1.86±0.38	75.40±0.04

Table 18: Tabular test results on HMDA with NaiveBayes as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	13.44±0.20	19.83±0.60	4.19±0.08	13.44±0.20	19.83±0.59	4.19±0.08	80.80±0.21
	WLMC	4.95±0.59	12.36±1.12	1.07±0.15	4.95±0.59	12.36±1.12	1.07±0.15	80.80±0.21
	Temp	9.47±0.20	17.90±0.58	2.03±0.05	9.47±0.20	17.90±0.58	2.03±0.05	80.80±0.21
	Platt	**3.72±0.19**	**11.27±0.25**	**0.98±0.07**	**3.72±0.19**	**11.27±0.25**	**0.98±0.07**	**81.70±0.18**
(PN,PU)	WLMC	3.47±0.91	**10.09±1.24**	0.80±0.30	3.83±1.13	**16.76±3.42**	1.02±0.47	80.83±0.22
	Temp	7.58±0.33	14.09±0.50	0.81±0.04	7.66±0.42	19.34±1.58	1.08±0.29	80.80±0.21
	Platt	**3.16±0.17**	13.11±0.54	**0.67±0.07**	**3.75±0.53**	20.01±2.77	**0.97±0.31**	**81.64±0.19**
(PN,Pconf)	WLMC	**3.15±0.40**	**11.47±0.92**	**0.91±0.12**	2.07±0.31	7.63±0.78	0.47±0.06	80.79±0.39
	Temp	8.25±0.26	13.23±0.37	1.28±0.03	8.54±0.29	14.56±0.36	0.75±0.07	80.80±0.21
	Platt	4.58±0.21	14.71±0.30	0.95±0.05	**1.37±0.17**	**5.73±0.30**	**0.36±0.03**	**80.81±0.18**
(PN,UU)	WLMC	4.97±0.75	**11.99±1.40**	1.08±0.14	5.04±0.78	**12.98±1.35**	1.13±0.21	80.80±0.21
	Temp	7.59±0.32	14.05±0.39	0.79±0.03	7.62±0.32	14.39±1.59	0.81±0.07	80.80±0.21
	Platt	**2.97±0.18**	12.91±0.42	**0.66±0.06**	**3.10±0.29**	14.09±1.69	**0.69±0.08**	**81.65±0.19**
(PU,PU)	ERM	19.68±5.01	29.76±2.78	9.86±8.04	19.74±5.02	31.53±3.24	9.99±8.05	76.11±1.08
	WLMC	2.24±0.77	**12.47±2.79**	**0.79±0.23**	2.45±0.94	**19.56±3.91**	**1.05±0.39**	76.57±1.11
	Temp	15.90±8.31	23.53±7.37	8.34±9.35	15.97±8.24	25.86±7.28	8.46±9.36	55.76±26.87
	Platt	**1.29±1.09**	14.66±1.66	1.46±1.21	**1.73±1.38**	19.85±4.26	1.70±1.14	**77.68±2.05**
(Pconf,Pconf)	ERM	15.62±0.18	25.92±0.74	3.11±0.12	14.97±0.15	25.29±0.62	2.76±0.07	75.19±0.14
	WLMC	4.50±0.49	**15.12±1.04**	0.52±0.05	4.38±0.43	13.03±3.26	0.56±0.04	77.08±0.20
	Temp	9.54±0.17	15.26±0.53	1.41±0.05	10.88±0.16	17.93±0.50	1.51±0.04	75.19±0.14
	Platt	**2.78±0.22**	16.11±0.27	**0.51±0.06**	**0.95±0.11**	**5.36±0.52**	**0.22±0.03**	**78.10±0.13**
(UU,UU)	ERM	14.93±3.73	20.64±4.52	5.92±4.68	15.04±3.72	22.34±3.90	6.04±4.70	80.12±1.77
	WLMC	4.52±1.39	**11.83±2.70**	1.10±0.19	4.64±1.37	**12.40±1.80**	1.14±0.17	80.12±1.78
	Temp	9.28±5.71	14.70±6.23	2.73±5.79	9.41±5.66	15.40±5.52	2.78±5.81	75.04±17.75
	Platt	**2.91±1.20**	12.95±1.57	**0.87±0.75**	**3.01±1.28**	13.99±1.88	**0.91±0.77**	**80.99±1.98**

Table 19: Tabular test results on HMDA with RandomForest as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	3.97±0.25	10.21±0.97	0.88±0.07	3.97±0.25	10.21±0.97	0.88±0.07	86.60±0.10
	WLMC	**3.09±0.17**	9.38±0.85	**0.39±0.04**	**3.09±0.17**	9.38±0.85	**0.39±0.04**	86.60±0.10
	Temp	7.70±14.11	13.35±13.00	2.36±5.40	7.70±14.11	13.36±13.00	2.36±5.40	79.28±23.14
	Platt	3.42±0.17	**6.67±0.82**	0.53±0.07	3.42±0.16	**6.67±0.82**	0.53±0.07	**86.62±0.10**
(PN,PU)	WLMC	3.11±0.46	9.25±0.99	0.48±0.20	3.71±0.80	20.54±4.43	0.82±0.33	86.60±0.10
	Temp	2.46±0.67	6.66±1.09	**0.34±0.16**	3.06±0.53	**18.34±2.54**	**0.69±0.23**	86.60±0.10
	Platt	**2.06±0.80**	**6.00±1.15**	0.37±0.14	**2.82±0.62**	18.93±2.26	0.70±0.28	**86.67±0.09**
(PN,Pconf)	WLMC	**3.97±0.25**	**10.21±0.97**	**0.88±0.07**	0.41±0.18	1.07±0.16	0.15±0.09	86.60±0.10
	Temp	4.42±0.18	10.61±0.76	0.99±0.04	0.24±0.09	0.93±0.18	0.05±0.02	86.60±0.10
	Platt	4.46±0.20	10.76±0.83	0.99±0.04	**0.23±0.08**	**0.93±0.17**	**0.05±0.02**	**86.60±0.10**
(PN,UU)	WLMC	2.97±0.26	9.20±0.78	0.37±0.02	2.99±0.34	11.22±2.08	0.42±0.07	86.60±0.10
	Temp	1.98±0.19	6.65±0.37	**0.22±0.03**	2.14±0.32	**10.33±0.71**	0.31±0.07	86.60±0.10
	Platt	**1.49±0.13**	**6.12±0.76**	0.23±0.04	**1.78±0.31**	10.96±1.62	**0.31±0.10**	**86.67±0.10**
(PU,PU)	ERM	17.60±2.91	31.69±5.65	6.18±2.14	17.31±2.84	35.62±6.59	6.06±2.24	76.98±1.95
	WLMC	**4.98±1.61**	19.12±3.14	**0.90±0.31**	**4.86±1.70**	25.52±5.59	**1.04±0.29**	**77.55±1.98**
	Temp	11.97±4.78	17.91±2.55	3.52±0.72	12.00±4.73	22.08±5.24	3.46±0.74	76.98±1.95
	Platt	5.90±2.76	**16.28±1.91**	2.29±0.76	6.01±2.82	**19.61±6.00**	2.49±1.04	76.33±3.99
(Pconf,Pconf)	ERM	12.06±0.13	22.07±0.56	2.20±0.10	7.00±0.11	9.60±0.20	1.22±0.04	77.62±0.22
	WLMC	8.09±0.27	17.13±0.72	0.93±0.09	3.41±0.25	8.48±0.37	0.53±0.09	81.83±0.42
	Temp	10.79±0.17	22.38±1.11	1.91±0.06	6.07±0.09	8.40±0.22	1.17±0.07	77.62±0.22
	Platt	**5.18±0.11**	**14.22±0.91**	**0.70±0.04**	**1.19±0.03**	**2.66±0.19**	**0.16±0.01**	**84.98±0.16**
(UU,UU)	ERM	19.80±2.61	24.91±3.62	6.87±2.09	19.82±2.73	29.13±5.46	6.88±2.22	79.62±2.07
	WLMC	**3.99±0.67**	10.51±0.50	**0.51±0.06**	**4.08±0.65**	11.65±1.46	**0.55±0.10**	80.36±1.28
	Temp	28.64±0.50	33.61±1.14	10.78±2.10	28.65±0.52	33.34±1.72	10.77±1.98	79.62±2.07
	Platt	7.17±1.18	**9.09±1.41**	2.02±0.37	7.23±1.04	**9.99±1.85**	2.03±0.34	**81.44±2.66**

Table 20: Tabular test results on HMDA with SVM as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	6.51±0.85	19.54±6.54	1.06±0.32	6.51±0.85	19.53±6.54	1.06±0.32	84.31±0.71
	WLMC	4.78±0.83	16.36±7.17	0.70±0.37	4.78±0.83	16.36±7.17	0.70±0.37	**84.33±0.72**
	Temp	**2.96±0.79**	17.17±6.58	**0.41±0.09**	**2.96±0.79**	17.15±6.61	**0.41±0.09**	84.31±0.71
	Platt	6.16±1.59	**14.26±8.66**	1.35±0.39	6.16±1.59	**14.26±8.66**	1.35±0.39	84.14±0.78
(PN,PU)	WLMC	4.67±1.25	16.53±7.39	0.64±0.13	4.78±1.18	**21.99±8.67**	0.82±0.24	84.34±0.71
	Temp	2.11±0.47	15.88±6.80	0.39±0.10	2.85±0.58	22.64±6.91	0.74±0.25	84.31±0.71
	Platt	**1.88±0.50**	**15.54±6.93**	**0.34±0.06**	**2.54±0.65**	22.60±7.16	**0.71±0.26**	**84.46±0.71**
(PN,Pconf)	WLMC	5.53±0.90	14.80±6.78	1.49±0.37	4.65±0.68	12.53±1.34	0.65±0.30	**84.33±0.69**
	Temp	**4.61±0.73**	**12.60±5.02**	1.34±0.11	1.76±0.49	8.79±3.02	0.59±0.12	84.31±0.71
	Platt	4.91±0.34	13.31±4.71	**1.30±0.09**	**1.42±0.39**	**8.48±2.82**	**0.44±0.14**	84.12±0.86
(PN,UU)	WLMC	4.79±0.86	15.73±7.06	0.80±0.46	4.95±0.88	16.03±4.93	0.85±0.44	84.32±0.70
	Temp	1.66±0.62	15.15±6.50	0.38±0.12	1.78±0.60	15.93±3.69	0.43±0.14	84.31±0.71
	Platt	**1.34±0.31**	**14.85±6.40**	**0.31±0.08**	**1.40±0.29**	**15.45±3.92**	**0.37±0.09**	**84.45±0.73**
(PU,PU)	ERM	12.80±1.65	25.97±8.18	3.56±0.79	12.46±1.95	29.21±9.20	3.52±0.93	82.12±1.39
	WLMC	4.13±1.02	**17.42±7.81**	0.93±0.39	4.28±1.26	**22.02±7.83**	1.23±0.48	**82.14±1.39**
	Temp	4.16±0.80	19.84±8.78	0.78±0.30	4.66±1.08	25.54±9.12	1.15±0.42	82.12±1.39
	Platt	**3.82±0.92**	19.61±8.35	**0.72±0.28**	**4.21±1.15**	24.84±8.91	**1.10±0.45**	81.99±1.51
(Pconf,Pconf)	ERM	9.03±2.11	31.61±6.96	2.28±1.08	8.07±2.32	19.12±6.47	2.16±1.17	76.69±0.70
	WLMC	4.71±0.64	27.91±7.18	**0.81±0.23**	3.36±0.43	18.23±7.48	**0.53±0.06**	**76.72±0.68**
	Temp	**3.48±1.53**	26.45±7.49	0.96±0.38	3.19±1.02	16.45±7.04	0.85±0.34	76.69±0.70
	Platt	4.33±1.18	**26.38±7.23**	1.05±0.47	**2.04±1.11**	**14.72±5.72**	0.84±0.44	76.53±0.84
(UU,UU)	ERM	2.43±0.85	9.67±3.90	0.42±0.12	2.73±0.88	14.06±3.96	0.44±0.13	85.48±0.58
	WLMC	2.33±0.78	9.52±3.85	0.37±0.09	2.63±0.83	13.80±3.47	0.39±0.11	85.48±0.58
	Temp	1.79±0.68	9.61±2.64	**0.33±0.06**	1.85±0.66	13.06±2.97	**0.35±0.04**	85.48±0.58
	Platt	**1.67±0.66**	**9.22±2.68**	0.37±0.12	**1.71±0.62**	**12.98±2.81**	0.37±0.10	**85.52±0.58**

Table 21: Tabular test results on HMDA with MLP as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	15.60±1.56	19.81±2.70	6.30±1.30	15.60±1.56	19.81±2.70	6.30±1.30	80.99±1.02
	WLMC	3.27±0.98	10.61±1.42	**0.63±0.23**	3.27±0.97	10.61±1.42	**0.63±0.23**	81.00±1.00
	Temp	12.78±1.95	16.78±3.02	4.64±1.19	12.78±1.95	16.78±3.02	4.64±1.19	80.99±1.02
	Platt	**2.91±1.15**	**8.73±1.73**	0.67±0.35	**2.91±1.15**	**8.73±1.73**	0.67±0.35	**81.04±0.80**
(PN,PU)	WLMC	**3.91±1.51**	10.23±2.52	1.25±0.71	**4.27±1.67**	**17.36±6.76**	1.44±0.86	81.00±1.00
	Temp	6.12±2.02	10.84±2.13	1.40±0.55	6.26±2.11	20.85±4.11	1.58±0.52	80.99±1.02
	Platt	4.53±0.93	**9.69±1.49**	**1.02±0.33**	4.71±0.99	20.71±4.25	**1.27±0.38**	**81.32±0.66**
(PN,Pconf)	WLMC	**3.08±1.61**	**10.59±2.03**	**0.83±0.56**	3.45±1.46	9.36±3.36	0.87±0.50	80.95±1.15
	Temp	6.23±2.34	11.09±2.46	1.82±0.59	5.17±2.95	8.29±3.63	1.15±0.61	80.99±1.02
	Platt	5.65±0.79	12.09±1.30	1.51±0.24	**3.05±0.82**	**5.09±0.71**	**0.80±0.24**	**81.10±0.63**
(PN,UU)	WLMC	**3.49±0.76**	10.90±1.76	**0.71±0.30**	**3.63±0.82**	12.29±2.08	**0.74±0.25**	81.00±1.00
	Temp	6.11±1.95	10.92±1.94	1.35±0.51	6.19±1.99	14.35±2.92	1.34±0.52	80.99±1.02
	Platt	4.52±0.93	**9.48±1.58**	0.95±0.33	4.74±0.99	**11.86±1.53**	0.95±0.38	**81.33±0.67**
(PU,PU)	ERM	17.29±6.78	26.77±10.09	8.16±4.40	17.36±6.67	29.50±8.80	8.30±4.39	77.16±1.39
	WLMC	3.69±1.97	13.77±2.46	**1.09±0.64**	4.31±2.14	**19.39±4.94**	**1.24±0.73**	77.17±1.39
	Temp	2.56±1.48	**12.18±2.49**	1.43±0.98	2.97±1.28	19.59±4.91	1.71±1.21	77.16±1.39
	Platt	**1.94±1.19**	12.57±2.37	1.25±0.91	**2.40±1.01**	19.73±4.29	1.59±1.17	**77.20±1.45**
(Pconf,Pconf)	ERM	18.40±2.62	27.25±4.92	8.89±2.59	16.53±2.07	19.43±3.21	8.20±2.29	77.79±1.08
	WLMC	5.80±1.70	15.92±3.28	1.67±0.99	4.46±1.07	10.43±2.81	1.26±0.61	77.78±1.07
	Temp	5.64±1.87	14.53±3.20	1.12±0.27	3.89±1.89	5.65±2.11	0.79±0.38	77.79±1.08
	Platt	**4.08±0.94**	**13.81±1.59**	**0.89±0.25**	**2.08±0.90**	**4.28±0.78**	**0.60±0.38**	**78.41±0.98**
(UU,UU)	ERM	7.68±4.35	13.05±4.56	2.01±1.97	7.82±4.31	14.75±4.27	2.12±2.07	81.90±1.01
	WLMC	3.85±1.26	9.43±1.92	0.58±0.33	4.01±1.20	11.29±3.10	0.64±0.33	82.15±1.10
	Temp	4.79±2.65	10.78±2.40	0.61±0.27	4.90±2.60	12.74±3.23	0.63±0.24	81.90±1.01
	Platt	**2.14±0.66**	**8.12±1.58**	**0.37±0.11**	**2.25±0.71**	**11.06±1.55**	**0.39±0.11**	**82.45±1.03**

Table 22: Tabular test results on MEPS with DecisionTree as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	8.74±0.72	30.22±2.60	1.82±0.22	8.74±0.72	30.21±2.60	1.82±0.22	83.82±0.20
	WLMC	4.25±1.30	**21.75±3.23**	**1.08±0.26**	4.25±1.30	**21.75±3.23**	**1.08±0.26**	**84.12±0.58**
	Temp	13.51±1.60	25.18±2.54	4.14±0.85	13.51±1.60	25.18±2.54	4.14±0.85	83.84±0.32
	Platt	**2.95±1.69**	23.28±1.06	2.07±0.52	**2.95±1.69**	23.28±1.06	2.07±0.52	83.60±0.21
(PN,PU)	WLMC	**4.58±1.08**	**23.46±3.14**	**1.17±0.39**	**6.00±1.36**	**27.94±9.41**	**1.54±0.38**	**84.04±0.77**
	Temp	26.95±2.15	32.63±2.72	13.45±1.89	27.53±2.92	35.46±5.43	13.41±1.83	83.84±0.32
	Platt	6.98±2.35	32.45±1.42	4.00±0.65	7.09±2.33	31.06±6.70	3.85±0.81	82.65±0.09
(PN,Pconf)	WLMC	**7.16±1.53**	**26.23±5.82**	**1.45±0.37**	8.10±2.28	**29.51±5.64**	**2.42±1.35**	**83.84±0.32**
	Temp	17.48±6.69	26.52±3.63	6.32±3.87	25.40±11.44	54.26±30.97	11.72±8.04	83.84±0.32
	Platt	7.42±1.91	36.12±3.02	2.43±0.76	**7.31±3.11**	34.65±3.34	4.40±2.48	82.82±0.42
(PN,UU)	WLMC	**4.96±1.76**	**20.21±2.37**	**1.26±0.46**	**5.49±1.96**	**24.84±3.88**	**1.39±0.46**	**84.23±0.50**
	Temp	26.44±1.40	31.73±1.80	12.76±1.41	26.56±1.11	34.29±4.01	12.74±1.38	83.84±0.32
	Platt	7.12±2.65	31.80±2.14	4.03±0.76	6.97±3.00	33.58±4.22	4.08±1.00	82.69±0.23
(PU,PU)	ERM	11.36±0.95	35.81±3.37	2.57±0.18	12.16±2.23	38.99±7.46	3.00±0.68	82.99±0.67
	WLMC	**5.29±1.55**	**19.70±5.44**	**1.36±0.59**	6.49±1.34	**24.41±7.78**	**1.61±0.63**	**83.34±1.01**
	Temp	28.06±1.69	34.86±2.64	14.98±3.00	28.15±2.42	37.31±2.44	14.91±2.99	82.99±0.66
	Platt	5.68±1.91	32.66±2.99	3.73±0.50	**5.19±2.76**	30.20±6.88	3.58±0.76	82.62±0.00
(Pconf,Pconf)	ERM	29.98±0.74	35.98±0.50	8.24±1.56	44.39±9.73	93.49±37.45	16.77±7.04	75.20±2.69
	WLMC	10.11±1.97	17.69±3.31	1.34±0.43	13.87±3.27	26.06±11.19	2.78±1.68	82.07±1.39
	Temp	25.41±2.78	34.94±1.91	4.96±1.39	34.41±10.48	65.10±29.22	9.03±6.88	75.20±2.69
	Platt	**3.96±1.31**	**16.22±1.92**	**0.80±0.41**	**4.29±0.96**	**14.61±2.72**	**1.56±1.20**	**85.22±0.41**
(UU,UU)	ERM	9.30±1.22	30.52±3.18	3.05±0.64	9.39±1.03	30.52±5.24	2.95±0.52	82.82±0.89
	WLMC	4.03±0.96	**20.56±3.29**	**0.90±0.20**	4.45±0.82	**25.18±4.91**	**1.00±0.18**	**84.03±0.50**
	Temp	27.42±2.43	33.32±2.84	12.64±2.89	27.47±2.27	36.03±3.83	12.71±3.13	82.82±0.89
	Platt	**2.08±0.91**	28.05±1.83	2.51±0.42	**2.30±0.83**	30.12±4.71	2.62±0.52	82.90±0.59

Table 23: Tabular test results on MEPS with LogisticRegression as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**2.04±0.27**	9.90±1.26	**0.54±0.12**	**2.04±0.27**	9.90±1.26	**0.54±0.12**	**86.28±0.26**
	WLMC	2.09±0.33	9.90±1.26	0.60±0.24	2.09±0.33	9.90±1.26	0.60±0.24	86.28±0.26
	Temp	3.43±0.38	**9.65±1.14**	1.31±0.09	3.43±0.37	**9.65±1.14**	1.31±0.09	86.28±0.26
	Platt	4.61±0.37	12.62±1.21	1.46±0.21	4.61±0.37	12.62±1.21	1.46±0.21	86.11±0.32
(PN,PU)	WLMC	**2.15±0.43**	**9.72±1.31**	**0.69±0.18**	**4.48±1.33**	24.15±4.66	**0.93±0.24**	86.17±0.47
	Temp	2.32±0.45	9.74±1.52	0.74±0.23	4.71±1.34	23.67±4.93	0.97±0.22	**86.28±0.26**
	Platt	2.70±0.62	11.61±1.58	0.77±0.24	4.89±1.69	**22.15±4.77**	0.98±0.28	86.12±0.28
(PN,Pconf)	WLMC	**2.56±0.48**	**9.93±1.18**	0.98±0.19	4.14±1.30	11.67±3.86	1.87±1.13	**86.28±0.26**
	Temp	3.24±0.99	11.29±2.14	**0.62±0.19**	**3.36±1.19**	11.64±1.74	**1.11±0.68**	86.28±0.26
	Platt	4.28±0.86	15.29±1.83	0.74±0.14	3.37±0.68	**11.59±1.29**	1.17±0.64	85.86±0.36
(PN,UU)	WLMC	**2.06±0.44**	10.16±1.47	**0.55±0.14**	**3.15±0.43**	19.65±3.05	**0.72±0.17**	**86.29±0.26**
	Temp	2.22±0.34	**9.89±1.33**	0.69±0.21	3.27±0.58	**18.64±3.84**	0.82±0.22	86.28±0.26
	Platt	2.48±0.57	11.40±1.82	0.73±0.25	3.48±0.79	19.92±3.98	0.84±0.25	86.10±0.28
(PU,PU)	ERM	3.61±0.91	12.89±2.84	1.19±0.35	3.64±0.82	16.50±4.63	1.06±0.27	84.27±0.42
	WLMC	3.14±0.81	12.82±2.44	**0.81±0.30**	3.34±1.41	**16.18±4.24**	1.02±0.23	84.34±0.40
	Temp	3.31±0.48	13.34±3.14	0.92±0.16	3.51±0.54	16.21±3.77	1.00±0.24	84.27±0.42
	Platt	**2.86±0.46**	**11.42±1.99**	0.90±0.18	**3.08±0.72**	16.69±4.07	**0.99±0.21**	**84.57±0.25**
(Pconf,Pconf)	ERM	35.46±1.56	42.60±1.13	10.20±2.31	51.25±12.24	106.75±50.48	18.77±9.25	56.74±7.59
	WLMC	8.85±1.95	15.76±3.35	1.93±0.97	15.12±7.00	32.20±15.29	4.39±3.25	81.20±2.30
	Temp	35.31±2.30	47.26±6.19	9.19±4.74	48.76±9.34	98.72±40.56	18.07±8.39	56.74±7.59
	Platt	**4.44±1.24**	**12.27±3.26**	**1.09±0.26**	**3.71±1.15**	**8.69±1.94**	**1.34±0.42**	**84.43±0.24**
(UU,UU)	ERM	3.25±0.48	11.42±2.15	1.10±0.18	3.71±0.47	**15.58±3.33**	1.10±0.15	84.48±0.46
	WLMC	**2.40±0.62**	12.04±2.29	**0.72±0.20**	**2.95±0.73**	16.09±3.48	**0.76±0.16**	84.36±0.50
	Temp	2.86±0.41	11.72±2.40	0.93±0.19	3.40±0.62	15.90±3.49	0.98±0.15	84.48±0.46
	Platt	2.73±0.75	**10.39±2.95**	0.93±0.20	3.30±0.98	16.68±2.19	0.97±0.17	**84.57±0.50**

Table 24: Tabular test results on MEPS with NaiveBayes as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	31.11±5.23	57.76±5.71	15.07±2.96	31.11±5.23	57.76±5.71	15.07±2.96	68.10±5.28
	WLMC	5.33±1.74	**13.99±3.94**	**1.17±0.27**	5.33±1.74	**13.99±3.94**	**1.17±0.27**	81.94±1.27
	Temp	30.53±5.21	56.96±5.79	13.57±2.32	30.53±5.21	56.96±5.79	13.57±2.32	68.10±5.28
	Platt	**2.72±0.66**	22.97±2.78	2.05±0.27	**2.72±0.66**	22.97±2.78	2.05±0.27	**82.62±0.00**
(PN,PU)	WLMC	6.12±1.67	**16.83±3.84**	**1.44±0.33**	6.91±1.57	20.23±4.71	1.78±0.48	82.33±1.87
	Temp	30.26±2.18	40.43±2.83	8.91±0.82	30.40±2.23	40.57±2.77	8.79±0.98	68.10±5.28
	Platt	**3.90±0.84**	20.80±2.22	1.82±0.19	**3.59±0.98**	**19.26±5.99**	**1.74±0.40**	**82.62±0.00**
(PN,Pconf)	WLMC	8.77±2.68	**18.32±4.76**	2.17±0.61	11.97±2.52	25.51±7.69	2.95±0.67	80.78±1.74
	Temp	28.82±3.65	45.80±5.24	8.55±0.87	37.69±6.99	67.57±14.65	11.95±4.79	68.10±5.28
	Platt	**4.09±0.81**	20.23±3.05	**1.77±0.28**	**2.57±1.21**	**11.38±3.64**	**1.16±0.26**	**82.62±0.00**
(PN,UU)	WLMC	6.02±1.98	**17.67±3.61**	**1.26±0.19**	6.24±1.64	**18.94±5.74**	**1.52±0.28**	82.24±1.49
	Temp	30.18±2.29	40.22±2.37	9.03±1.00	30.05±2.42	40.56±3.83	9.26±0.86	68.10±5.28
	Platt	**4.23±1.22**	21.24±3.07	1.85±0.27	**4.10±1.43**	23.52±5.77	2.08±0.39	**82.62±0.00**
(PU,PU)	ERM	36.49±9.63	60.50±16.80	22.00±6.90	36.97±10.26	59.37±17.15	22.13±6.64	48.21±23.16
	WLMC	5.68±2.27	**16.94±4.31**	**1.41±0.54**	6.43±2.85	**18.40±3.90**	**1.70±0.57**	81.84±2.13
	Temp	31.33±3.59	41.14±4.05	15.19±6.69	31.36±3.57	41.75±3.67	14.93±6.52	67.78±13.73
	Platt	**2.32±1.42**	27.70±7.74	2.52±0.79	**2.33±1.44**	26.20±9.69	2.48±0.98	**82.62±0.00**
(Pconf,Pconf)	ERM	20.82±1.05	42.37±2.57	8.72±0.57	23.58±1.37	62.92±4.23	8.28±1.04	77.79±0.93
	WLMC	9.27±3.62	19.34±7.23	2.06±0.78	11.70±4.14	28.57±12.77	2.48±0.66	82.28±1.80
	Temp	20.42±1.86	36.92±2.87	4.32±0.61	25.47±4.45	50.95±6.27	5.63±2.26	77.79±0.93
	Platt	**3.01±0.86**	**14.12±1.88**	**1.03±0.14**	**2.40±0.78**	**6.83±1.91**	**1.12±0.45**	**83.11±0.55**
(UU,UU)	ERM	32.62±0.00	44.72±0.00	24.47±0.00	32.62±0.00	45.73±2.29	24.33±0.85	17.38±0.00
	WLMC	3.03±1.67	**16.25±6.12**	**1.17±0.75**	3.59±1.55	**20.07±4.50**	**1.38±0.78**	**82.78±0.40**
	Temp	32.62±0.00	44.72±0.00	24.47±0.00	32.62±0.00	45.73±2.29	24.33±0.85	82.62±0.00
	Platt	**0.81±0.58**	36.93±0.58	3.51±0.13	**0.81±0.58**	38.47±3.22	3.77±0.26	82.62±0.00

Table 25: Tabular test results on MEPS with RandomForest as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	1.76±0.41	**8.94±1.39**	**0.41±0.07**	1.76±0.41	**8.94±1.39**	**0.41±0.07**	**86.32±0.27**
	WLMC	**1.74±0.39**	9.05±1.46	0.45±0.18	**1.74±0.39**	9.06±1.46	0.45±0.18	86.32±0.27
	Temp	3.82±0.50	9.44±1.16	0.88±0.25	3.82±0.50	9.44±1.16	0.88±0.25	86.32±0.27
	Platt	2.63±0.47	10.62±0.97	0.83±0.23	2.63±0.47	10.62±0.97	0.83±0.23	86.27±0.20
(PN,PU)	WLMC	**2.14±0.54**	10.18±2.36	0.59±0.24	**4.01±1.21**	21.99±4.17	1.06±0.23	86.26±0.33
	Temp	2.18±0.50	**9.16±1.44**	0.54±0.08	4.47±1.03	**19.83±3.25**	0.88±0.21	**86.32±0.27**
	Platt	2.68±0.86	11.55±3.05	**0.52±0.12**	4.34±1.25	21.86±4.31	**0.86±0.19**	85.98±0.56
(PN,Pconf)	WLMC	2.03±0.58	**9.01±1.33**	0.62±0.29	1.80±1.42	6.58±5.99	1.28±1.37	86.32±0.27
	Temp	1.77±0.31	9.19±1.29	0.39±0.09	0.82±0.22	2.46±0.52	**0.21±0.05**	86.32±0.27
	Platt	**1.76±0.27**	9.43±1.58	**0.37±0.07**	**0.80±0.22**	**2.29±0.63**	0.21±0.05	**86.34±0.24**
(PN,UU)	WLMC	**1.98±0.79**	**9.22±1.77**	0.51±0.19	**2.97±1.04**	**15.99±3.32**	0.70±0.22	86.20±0.27
	Temp	2.13±0.49	9.27±1.62	0.54±0.19	3.09±0.73	16.85±3.80	0.73±0.22	**86.32±0.27**
	Platt	2.35±1.00	10.54±2.79	**0.46±0.13**	3.21±0.72	16.84±3.24	**0.66±0.19**	85.90±0.56
(PU,PU)	ERM	17.85±13.52	32.36±15.42	7.20±10.27	18.39±13.62	37.24±13.87	7.53±10.10	77.30±11.63
	WLMC	4.67±1.19	**15.77±4.64**	**1.20±0.57**	5.78±1.33	**20.74±1.93**	**1.40±0.65**	83.15±2.30
	Temp	26.07±4.98	34.63±6.24	9.03±4.05	26.44±5.10	37.34±5.87	8.90±3.95	77.30±11.63
	Platt	**3.70±1.85**	26.55±5.57	2.14±0.94	**4.08±2.43**	25.43±5.29	2.15±0.87	**83.65±1.14**
(Pconf,Pconf)	ERM	30.75±0.49	36.49±0.58	12.71±1.38	47.04±11.39	106.32±48.48	22.33±10.83	73.96±1.04
	WLMC	7.69±2.14	**15.28±3.32**	1.49±0.61	12.68±5.18	25.52±12.37	3.51±2.20	83.02±1.83
	Temp	24.44±2.25	35.58±2.75	3.11±0.71	33.47±9.33	67.57±26.99	9.24±7.22	73.96±1.04
	Platt	**4.08±0.89**	15.41±2.41	**0.72±0.27**	**3.65±1.18**	**12.34±3.68**	**1.49±1.19**	**85.27±0.19**
(UU,UU)	ERM	40.75±1.79	55.44±1.80	15.54±2.32	40.78±1.60	58.13±5.38	15.51±2.43	59.20±1.85
	WLMC	7.82±1.01	**20.22±3.13**	**1.69±0.31**	8.06±1.31	**24.09±5.03**	**1.80±0.33**	79.61±1.28
	Temp	33.66±0.14	44.46±0.28	15.80±0.64	33.66±0.12	45.55±2.32	15.74±0.84	59.20±1.85
	Platt	**1.54±1.01**	23.11±1.93	1.89±0.22	**1.56±0.88**	24.89±5.61	2.13±0.27	**82.62±0.00**

Table 26: Tabular test results on MEPS with SVM as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	11.38±2.53	28.32±7.50	2.55±1.19	11.38±2.53	28.32±7.50	2.55±1.19	83.07±2.17
	WLMC	5.02±1.40	**18.45±4.73**	**0.93±0.33**	5.02±1.40	**18.45±4.73**	**0.93±0.33**	84.30±1.19
	Temp	8.59±3.17	23.37±8.09	1.49±0.84	8.59±3.17	23.37±8.09	1.49±0.84	83.07±2.17
	Platt	**2.03±0.66**	21.15±6.43	1.45±0.81	**2.03±0.66**	21.15±6.43	1.45±0.81	**84.52±1.27**
(PN,PU)	WLMC	5.26±1.04	**19.51±5.52**	**1.21±0.74**	6.08±1.61	**24.44±6.45**	**1.48±0.79**	**84.24±1.40**
	Temp	10.47±2.71	22.09±4.76	2.07±0.53	11.06±2.91	27.96±3.66	2.13±0.71	83.07±2.17
	Platt	**3.28±1.09**	23.34±6.51	1.58±0.84	**4.43±1.25**	27.82±8.14	1.66±0.78	84.21±0.97
(PN,Pconf)	WLMC	8.18±3.49	**21.96±7.45**	1.58±0.64	10.63±5.37	29.59±6.93	2.12±0.44	83.48±2.06
	Temp	9.02±2.96	22.63±7.02	1.41±0.54	12.65±3.86	29.66±5.27	2.11±0.73	83.07±2.17
	Platt	**5.46±2.17**	26.88±8.34	**1.36±0.70**	**4.76±1.91**	**23.15±8.54**	**1.66±0.92**	**84.27±0.90**
(PN,UU)	WLMC	4.72±1.10	**18.26±4.43**	**1.02±0.30**	5.53±1.47	**26.02±6.24**	**1.21±0.45**	84.18±1.25
	Temp	10.34±2.15	21.62±5.01	1.98±0.54	10.59±2.21	26.53±5.52	2.06±0.58	83.07±2.17
	Platt	**3.06±1.01**	23.19±6.66	1.57±0.86	**3.92±1.40**	28.17±6.29	1.68±0.95	**84.31±0.87**
(PU,PU)	ERM	17.56±1.48	39.98±6.64	5.19±0.80	18.19±1.62	42.43±7.52	5.37±1.40	81.12±1.69
	WLMC	5.81±1.23	**22.03±7.65**	**1.31±0.34**	6.62±1.56	27.49±8.48	**1.63±0.40**	83.23±0.81
	Temp	9.98±3.07	23.43±5.31	3.05±1.11	9.89±3.33	28.11±5.30	3.18±1.34	81.12±1.69
	Platt	**2.16±0.96**	26.85±6.02	1.73±0.76	**3.00±1.74**	**26.34±7.80**	1.90±0.72	**83.60±1.11**
(Pconf,Pconf)	ERM	39.12±5.48	65.96±13.65	18.71±5.16	54.88±23.52	126.15±87.86	31.30±20.86	57.80±4.85
	WLMC	11.19±3.00	**25.81±5.27**	**2.32±0.93**	14.45±3.84	34.82±9.94	**4.52±2.67**	80.04±2.37
	Temp	31.92±2.58	46.20±3.36	10.90±4.96	46.46±13.23	100.96±49.07	19.75±11.57	57.80±4.85
	Platt	**5.71±3.23**	35.03±8.14	2.58±1.08	**7.20±3.15**	**34.76±8.53**	4.75±3.33	**82.80±0.56**
(UU,UU)	ERM	13.54±1.78	33.62±5.97	3.60±0.53	13.74±1.77	37.62±8.18	3.66±0.54	82.27±1.70
	WLMC	4.97±1.09	**19.39±3.13**	**1.00±0.38**	5.51±1.06	**25.86±4.93**	**1.14±0.39**	83.31±1.61
	Temp	10.78±2.62	21.44±3.97	2.24±0.66	10.92±2.45	25.95±4.96	2.31±0.70	82.27±1.70
	Platt	**3.31±0.72**	26.31±6.21	1.87±0.81	**3.55±0.95**	30.04±6.70	1.94±0.88	**83.53±0.98**

Table 27: Tabular test results on MEPS with MLP as the base predictor. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over ten seeds; accuracy is reported in percent. WLMC uses step size 0.05 and at most 50 updates. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	**2.46±0.59**	10.02±1.21	**0.74±0.30**	**2.46±0.59**	10.02±1.21	**0.74±0.30**	**86.50±0.35**
	WLMC	2.51±0.70	10.14±1.24	0.85±0.40	2.51±0.70	10.14±1.24	0.85±0.40	86.50±0.35
	Temp	3.77±0.65	**9.94±0.75**	1.33±0.21	3.77±0.65	**9.94±0.76**	1.33±0.21	86.50±0.35
	Platt	4.27±0.46	11.29±1.25	1.35±0.26	4.27±0.46	11.29±1.25	1.35±0.26	86.41±0.26
(PN,PU)	WLMC	**2.32±0.71**	9.71±1.48	0.80±0.46	4.60±1.09	21.20±3.64	0.95±0.30	86.50±0.35
	Temp	2.57±0.51	**9.56±1.27**	**0.68±0.23**	4.65±1.14	20.98±3.95	**0.92±0.19**	**86.50±0.35**
	Platt	2.69±0.44	10.21±2.10	0.69±0.19	**4.59±1.11**	**20.33±4.48**	0.93±0.22	86.38±0.28
(PN,Pconf)	WLMC	**2.46±0.59**	**9.89±1.45**	0.89±0.27	4.29±1.58	11.71±6.49	2.28±1.60	**86.50±0.35**
	Temp	3.17±0.79	10.84±1.73	**0.70±0.26**	3.34±0.80	9.53±2.24	**1.16±0.62**	86.50±0.35
	Platt	3.84±0.73	12.47±2.14	0.79±0.24	**3.27±0.89**	**8.91±2.08**	1.18±0.61	86.32±0.32
(PN,UU)	WLMC	2.70±0.77	10.42±1.74	0.92±0.36	3.65±0.63	**16.33±2.94**	0.95±0.34	**86.51±0.33**
	Temp	**2.54±0.57**	**9.90±1.47**	**0.67±0.25**	**3.39±0.71**	17.31±2.81	**0.76±0.18**	86.50±0.35
	Platt	2.66±0.62	10.36±2.09	0.67±0.25	3.51±0.73	16.85±3.61	0.78±0.20	86.39±0.27
(PU,PU)	ERM	2.90±0.84	11.14±2.70	0.88±0.26	4.58±1.17	19.96±3.28	1.03±0.22	**85.71±0.17**
	WLMC	2.98±1.03	10.94±2.98	1.03±0.48	4.27±1.21	**18.79±2.96**	1.29±0.49	85.48±0.62
	Temp	2.45±0.92	**10.74±2.87**	0.83±0.36	4.24±0.93	20.37±3.74	1.05±0.31	85.71±0.17
	Platt	**2.28±0.54**	10.96±1.54	**0.81±0.29**	**3.96±1.26**	18.82±4.93	**1.00±0.21**	85.33±0.28
(Pconf,Pconf)	ERM	26.70±1.07	32.90±1.35	12.30±0.82	41.24±10.11	94.39±43.41	21.88±9.92	83.32±0.90
	WLMC	7.67±1.27	**14.28±1.66**	1.24±0.45	11.75±5.07	22.65±10.46	2.80±2.22	84.48±0.81
	Temp	11.38±2.87	23.14±4.65	2.13±0.48	16.89±4.87	41.35±8.25	4.86±3.42	83.32±0.90
	Platt	**4.54±1.31**	15.19±2.68	**0.94±0.34**	**4.86±1.16**	**14.51±3.21**	**1.87±1.10**	**85.96±0.22**
(UU,UU)	ERM	75.05±2.42	88.14±0.93	36.51±3.21	74.89±2.53	89.36±3.35	35.66±4.08	17.61±0.63
	WLMC	12.34±2.17	**25.67±4.63**	2.78±0.42	12.49±2.23	**27.38±5.07**	2.91±0.54	75.54±2.95
	Temp	34.44±0.31	45.09±0.13	25.47±0.38	34.43±0.32	46.38±2.11	25.29±1.01	17.61±0.63
	Platt	**2.75±0.76**	27.10±3.86	**2.08±0.43**	**3.07±0.89**	29.15±4.89	**2.31±0.43**	**82.62±0.00**

Table 28: CelebA-ViT test results. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation; accuracy is reported in percent. Bold marks the single best mean within each $(\text{base}, \text{post})$ block; variance is not used for bolding.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	1.68±0.60	6.05±0.40	0.60±0.14	1.68±0.60	6.05±0.40	0.60±0.14	92.11±0.27
	WLMC	**1.62±0.68**	**5.81±0.48**	**0.45±0.11**	**1.62±0.68**	**5.81±0.48**	**0.45±0.11**	92.11±0.27
	Temp	2.72±0.23	6.75±0.68	0.90±0.04	2.72±0.23	6.75±0.68	0.90±0.04	92.11±0.27
	Platt	2.00±0.22	5.81±0.39	1.00±0.13	2.00±0.22	5.81±0.39	1.00±0.13	**92.17±0.21**
(PN,PU)	ERM	1.68±0.60	6.05±0.40	0.60±0.14	1.86±0.42	7.46±0.89	0.66±0.18	92.11±0.27
	WLMC	1.65±0.64	**5.69±0.50**	0.51±0.07	1.83±0.46	**7.09±0.88**	0.56±0.06	92.11±0.27
	Temp	1.53±0.51	5.94±0.53	0.54±0.04	1.66±0.40	7.45±1.02	0.56±0.07	92.11±0.27
	Platt	**0.61±0.19**	5.81±0.45	**0.49±0.05**	**0.80±0.33**	7.46±0.73	**0.49±0.05**	**92.18±0.24**
(PN,UU)	ERM	1.68±0.60	6.05±0.40	0.60±0.14	1.75±0.53	6.53±0.98	0.62±0.14	92.11±0.27
	WLMC	1.62±0.68	**5.81±0.48**	**0.45±0.11**	1.67±0.59	6.29±0.77	**0.47±0.10**	92.11±0.27
	Temp	1.49±0.46	5.90±0.45	0.52±0.06	1.52±0.41	6.06±1.32	0.54±0.06	92.11±0.27
	Platt	**0.66±0.20**	5.90±0.49	0.47±0.02	**0.78±0.15**	**5.95±0.11**	0.49±0.02	**92.17±0.24**
(PU,PU)	ERM	1.37±0.49	7.37±1.13	0.66±0.13	1.54±0.70	8.05±0.90	0.71±0.11	90.97±0.79
	WLMC	1.23±0.11	7.14±1.03	**0.53±0.04**	1.39±0.28	**7.94±1.01**	**0.55±0.06**	**91.04±0.65**
	Temp	1.29±0.36	**7.12±1.03**	0.65±0.14	1.42±0.47	8.41±1.15	0.68±0.14	90.97±0.79
	Platt	**1.05±0.28**	7.27±1.05	0.62±0.12	**1.20±0.34**	7.98±1.06	0.65±0.11	90.95±0.75
(UU,UU)	ERM	1.39±0.41	7.14±0.72	0.52±0.07	1.41±0.38	7.77±0.88	0.52±0.08	91.49±0.28
	WLMC	1.38±0.43	6.91±1.10	**0.47±0.11**	1.41±0.37	7.81±0.73	**0.47±0.12**	91.49±0.28
	Temp	1.32±0.32	6.97±0.95	0.49±0.07	1.37±0.33	7.68±0.90	0.51±0.06	91.49±0.28
	Platt	**0.65±0.17**	**5.99±0.45**	0.53±0.09	**0.78±0.24**	**7.41±1.18**	0.53±0.06	**91.65±0.34**

L Large-model experimental supplements

This section provides compact test summaries for the CelebA-ViT image model, the CelebA-ImageResNet companion image model, and the CivilComments-BERT text model. The protocol, metric, weak-data construction, and hyperparameter definitions are given in Appendices G–I; Tables 28, 29, and 30 focus on test-split results.

Large-model result summary. The large-model results show that weak post-processing remains effective for both image and text classifiers, but the relative ranking of the post-processing methods is more mixed than in the tabular experiments. For CelebA-ViT and CelebA-ImageResNet, WLMC often gives strong improvements on subgroup-sensitive calibration metrics such as MC and maxECE, while Platt scaling and temperature scaling are also competitive in several settings. For CivilComments-BERT, the global post-processing baselines are particularly strong: Platt scaling and temperature scaling frequently achieve large gains and can match or outperform WLMC on some reported metrics. Thus, these experiments show that the proposed WSL framework supports effective weak post-processing across modern image and text models, rather than as showing uniform dominance of WLMC over simpler calibration methods. Overall, WLMC is most useful for subgroup-sensitive correction, while Platt and temperature scaling remain strong baselines, especially for the CivilComments-BERT task.

L.1 CelebA-ViT

Table 28 reports the test-split values underlying the CelebA-ViT entries in Figure 4.

L.2 CelebA-ImageResNet

Table 29 reports the compact test-split results for the companion ResNet image model.

L.3 CivilComments-BERT

Table 30 reports the test-split values underlying the CivilComments-BERT entries in Figure 4.

Table 29: CelebA-ImageResNet test results. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation over five seeds; accuracy is reported in percent. Bold marks the best mean within each $(\text{base}, \text{post})$ block; ties are all bolded.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN, PN)	ERM	**1.01±0.49**	**3.40±1.00**	**0.33±0.08**	**1.01±0.49**	**3.40±1.00**	**0.33±0.08**	94.38±0.15
	WLMC	**1.01±0.49**	**3.40±1.00**	**0.33±0.08**	**1.01±0.49**	**3.40±1.00**	**0.33±0.08**	94.38±0.15
	Temp	1.54±0.69	4.80±1.02	0.43±0.07	1.54±0.69	4.80±1.02	0.43±0.07	94.38±0.15
	Platt	1.48±0.16	3.46±0.49	0.60±0.09	1.48±0.16	3.46±0.49	0.60±0.09	**94.42±0.05**
(PN, PU)	ERM	1.01±0.49	**3.40±1.00**	0.33±0.08	1.16±0.37	4.94±1.32	0.40±0.09	94.38±0.15
	WLMC	1.01±0.49	**3.40±1.00**	0.33±0.08	1.16±0.37	4.94±1.32	0.40±0.09	94.38±0.15
	Temp	0.90±0.74	3.75±1.11	0.28±0.06	1.10±0.55	**4.90±1.20**	0.32±0.07	94.38±0.15
	Platt	**0.43±0.09**	3.44±0.29	**0.25±0.03**	**0.80±0.07**	4.96±1.07	**0.29±0.07**	**94.42±0.07**
(PN, UU)	ERM	1.01±0.49	**3.40±1.00**	0.33±0.08	1.12±0.47	**3.74±0.83**	0.36±0.09	94.38±0.15
	WLMC	1.01±0.49	**3.40±1.00**	0.33±0.08	1.12±0.47	**3.74±0.83**	0.36±0.09	94.38±0.15
	Temp	0.87±0.77	3.87±1.08	0.26±0.08	1.07±0.67	4.29±1.15	0.27±0.08	94.38±0.15
	Platt	**0.39±0.17**	3.52±0.43	**0.24±0.06**	**0.69±0.06**	3.78±0.66	**0.25±0.07**	**94.41±0.06**
(PU, PU)	ERM	1.33±0.51	**5.21±0.31**	0.47±0.18	1.47±0.52	6.38±0.91	0.49±0.14	**93.23±0.16**
	WLMC	1.22±0.43	5.23±0.35	0.40±0.10	1.35±0.45	6.30±0.71	0.42±0.11	**93.23±0.16**
	Temp	1.05±0.43	5.48±0.41	0.37±0.06	1.14±0.49	**5.84±0.75**	0.39±0.15	**93.23±0.16**
	Platt	**0.81±0.18**	6.15±1.00	**0.31±0.05**	**0.93±0.22**	6.28±0.92	**0.31±0.07**	93.22±0.18
(UU, UU)	ERM	1.46±0.68	4.97±1.04	0.56±0.22	1.51±0.56	**5.29±1.01**	0.50±0.17	**93.50±0.24**
	WLMC	1.29±0.48	**4.82±0.89**	0.39±0.08	1.35±0.39	5.53±1.23	0.38±0.08	**93.50±0.24**
	Temp	1.11±0.81	5.83±1.36	**0.35±0.07**	1.21±0.67	6.33±1.55	**0.36±0.07**	**93.50±0.24**
	Platt	**0.82±0.34**	5.33±0.66	**0.35±0.06**	**1.06±0.38**	5.89±0.85	**0.36±0.07**	**93.50±0.23**

Table 30: CivilComments-BERT test results. Calibration metrics are reported as $10^2\times$ mean $\pm$ standard deviation; accuracy is reported in percent. Bold marks the single best reported mean within each $(\text{base}, \text{post})$ block.

(Base,Post)	Post-processing	oracle estimate ($\times 10^{-2}$)			weak estimate ($\times 10^{-2}$)			Acc. (%)
		ECE	maxECE	MC	ECE	maxECE	MC	
(PN,PN)	ERM	1.64±0.92	9.80±9.82	0.87±0.92	1.64±0.92	9.79±9.82	0.87±0.92	**90.96±1.30**
	WLMC	1.35±0.56	9.16±8.94	**0.61±0.43**	1.35±0.56	9.16±8.94	**0.61±0.43**	90.96±1.30
	Temp	2.48±0.77	**8.64±7.63**	1.52±1.24	2.48±0.77	**8.64±7.62**	1.52±1.24	90.96±1.30
	Platt	**1.31±0.88**	10.83±7.74	0.65±0.35	**1.31±0.88**	10.83±7.74	0.65±0.35	90.94±1.29
(PN,PU)	ERM	1.64±0.92	9.80±9.82	0.87±0.92	1.64±0.92	**9.79±9.82**	0.87±0.92	**90.96±1.30**
	WLMC	1.27±0.59	9.26±8.93	0.58±0.45	1.35±0.59	11.72±7.91	0.61±0.39	90.96±1.30
	Temp	0.72±0.57	8.76±8.90	0.40±0.46	0.79±0.57	11.42±7.89	0.44±0.46	90.96±1.30
	Platt	**0.59±0.48**	**8.10±9.34**	**0.37±0.47**	**0.75±0.47**	10.74±8.20	**0.43±0.47**	90.89±1.26
(PN,UU)	ERM	1.64±0.92	9.80±9.82	0.87±0.92	1.64±0.92	9.79±9.82	0.87±0.92	**90.96±1.30**
	WLMC	1.35±0.55	9.19±8.93	0.61±0.43	1.38±0.57	10.04±9.06	0.60±0.42	90.96±1.30
	Temp	0.71±0.57	8.83±8.87	0.41±0.46	0.74±0.60	10.22±8.80	0.43±0.47	90.96±1.30
	Platt	**0.57±0.47**	**8.02±9.44**	**0.37±0.47**	**0.61±0.48**	**9.17±9.35**	**0.39±0.49**	90.88±1.26
(PU,PU)	ERM	2.79±1.78	10.78±5.42	1.24±1.05	2.79±1.81	11.56±4.00	1.21±1.06	90.81±0.09
	WLMC	2.27±1.29	9.26±3.69	0.70±0.48	2.24±1.34	10.01±2.01	0.65±0.46	90.81±0.09
	Temp	1.62±0.80	8.78±3.93	0.47±0.26	1.61±0.86	10.01±2.60	0.47±0.26	90.81±0.09
	Platt	**1.06±0.16**	**5.88±1.66**	**0.30±0.10**	**1.00±0.24**	**8.31±1.63**	**0.32±0.10**	**90.90±0.11**
(UU,UU)	ERM	1.56±0.53	6.21±1.12	0.54±0.21	1.57±0.48	7.14±0.80	0.55±0.21	**91.22±0.16**
	WLMC	1.46±0.49	6.00±1.22	0.47±0.12	1.47±0.44	6.94±0.91	0.47±0.12	91.22±0.16
	Temp	1.22±0.66	5.69±1.16	0.27±0.14	1.22±0.62	**6.68±0.94**	0.28±0.15	91.22±0.16
	Platt	**1.08±0.55**	**5.52±0.82**	**0.25±0.11**	**1.09±0.55**	6.70±0.76	**0.25±0.11**	91.20±0.13

L.4 Supplementary notes on the large-model improvement heatmap

Figure 8 reproduces the two-panel improvement heatmap used in the main-text large-model comparison. The panels use the same settings as the CelebA-ViT and CivilComments-BERT compact result tables, Tables 28 and 30. For each dataset, setting, and metric $m \in \{\text{ECE}, \text{MC}\}$, each cell in Figure 8 displays the best reported test mean among WLMC, temperature scaling, and Platt scaling for that metric. This descriptive comparison is distinct from validation-based selection of a fitted

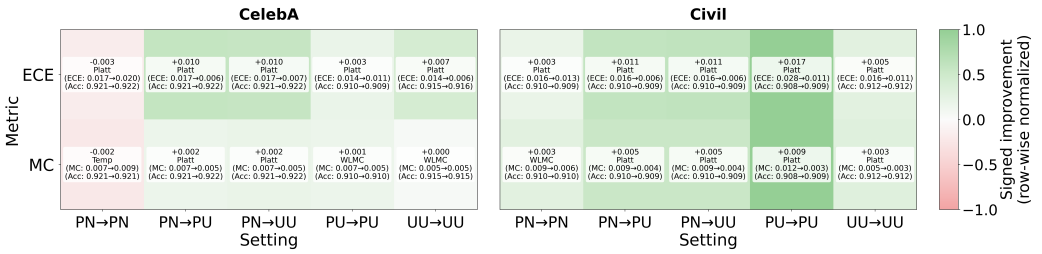


Figure 8: Supplementary view of the large-model improvement heatmap. For each ECE or MC cell, the displayed method has the best reported test mean among WLMC, temperature scaling, and Platt scaling for that metric. The cell reports the signed improvement, the selected method, the calibration-metric before/after values, and the accuracy before/after values. Colors are row-wise normalized over both panels.

predictor. The first line in each cell is the signed improvement

$$\Delta m = m_{\mathrm{ERM}} - m_{\mathrm{post}},$$

so positive values correspond to a reduction in calibration error. The remaining lines in each cell of Figure 8 list the selected method, the metric transition $m_{\mathrm{ERM}} \to m_{\mathrm{post}}$, and the corresponding accuracy transition $\mathrm{Acc}_{\mathrm{ERM}} \to \mathrm{Acc}_{\mathrm{post}}$ for the same selected method. Accuracy is therefore reported as contextual information for the calibration-improving method rather than as a separate heatmap row.

The colors summarize signed improvement after row-wise normalization over both panels: the ECE row is normalized by the largest absolute ECE improvement across CelebA and CivilComments, and the MC row is normalized analogously. This makes the colors comparable between CelebA and CivilComments within the same metric row, while the exact numerical entries should be used when comparing ECE against MC. A near-white cell means little change relative to the largest change in that row, green indicates improvement, and red indicates degradation.

Figures 9–10 provide supplementary CelebA-ViT diagnostics. The MC scatter uses the additional setting $\eta = 0.02$ and update cap 100, as indicated by its plotted run label; it is not the $\eta = 0.04$, cap-50 setting in Table 3. Per-group ECE CDFs show how post-processing shifts the group-wise ECE distribution, while the MC before/after scatter tracks changes in the supremum-style group–bin violation.

Subgroup-family diagnostics. These plots also partially address the sensitivity of the conclusions to the evaluated subgroup family. They are not a full sweep over alternative families: all panels use the fixed finite CelebA subgroup collection described in Appendix G and the same ten prediction bins as the main experiments. However, the CDF view is useful because it shows whether post-processing improves calibration broadly across the evaluated subgroups or only changes a small number of extreme cases. A leftward shift of the per-group ECE CDF indicates improvement throughout the subgroup family, while a change only in the upper tail would suggest that the method is mainly affecting the worst groups. The MC scatter complements this distributional view by tracking the supremum-type group–bin violation. Thus, Figures 9–10 should be read as a subgroup-family diagnostic for the current interpretable group collection. A stronger robustness check would additionally sweep the number of prediction bins or add and remove subgroup intersections; the current CDF/scatter analysis provides evidence over the reported family but does not replace that alternative-family sensitivity study.

M Impact Statement

This paper studies multicalibration under weak supervision, where clean input–label pairs are unavailable or costly but weaker observations such as positive-unlabeled, unlabeled-unlabeled, or positive-confidence data are available. Its potential positive societal impact is to make calibration auditing and post-hoc recalibration more feasible in settings where reliable probability estimates are important but fully labeled calibration data are limited. By providing corrected multicalibration

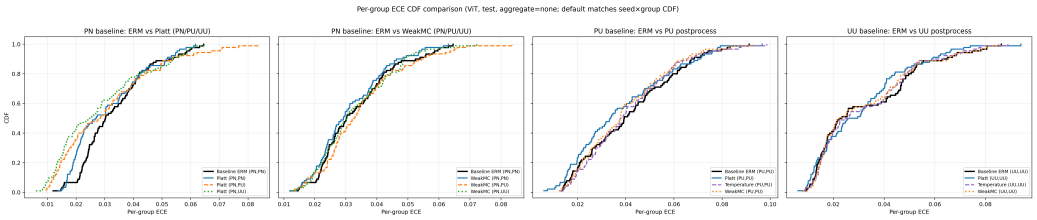


Figure 9: Per-group ECE CDFs on CelebA-ViT (test split). A leftward shift indicates improvement of the per-group ECE distribution. Platt scaling is especially effective for ECE-like metrics, whereas WLMC is more conservative in ECE terms.

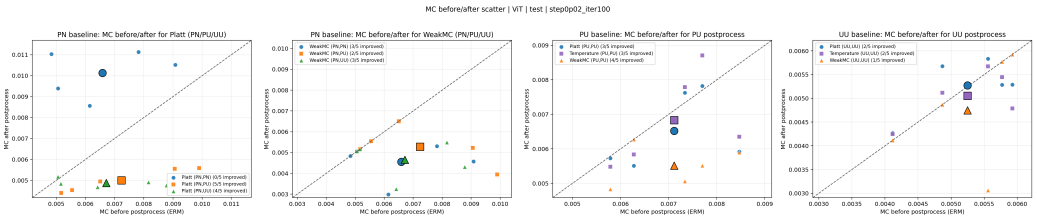


Figure 10: MC before/after scatter plots on CelebA-ViT (test split; $\eta = 0.02$, update cap 100). Points below the diagonal correspond to reduced MC after post-processing correction. WLMC is more directly aligned with MC reduction, while Platt and temperature can show mixed seed-level behavior depending on the regime.

estimators and weak-label post-processing methods with finite-sample guarantees, the work may help practitioners evaluate subgroup- and score-dependent calibration under weaker data-access conditions.

The work also has limitations that are relevant to its societal interpretation. The guarantees rely on assumptions about the weak-supervision mechanism, such as class priors, mixture proportions, confidence values, and the correctness of the risk-rewrite model. If these assumptions are misspecified, the resulting calibration estimates or post-processing updates may be misleading. A small weakly supervised multicalibration estimate should therefore not be interpreted as a complete deployment guarantee: it does not ensure fairness, representativeness, robustness, causal validity, or suitability for downstream decisions. A possible negative impact would be over-reliance on weak-label calibration estimates as a substitute for clean-label validation, dataset documentation, subgroup analysis, or application-specific auditing.

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 claims in the abstract and Section 1 match our theoretical and numerical claims in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The theorem statements specify their assumptions. Section 7 discusses the exact weak-observation model and finite-class restriction of the fixed-sample guarantee; Appendix I.3 distinguishes the experimental implementation from the theorem.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical statements explicitly state the assumptions. Complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer:[Yes]

Justification: The appendix reports the datasets, weak-observation constructions, model classes, data splits, seeds, calibration fractions, post-processing methods, evaluation metrics, and hyperparameters used in the experiments. The supplemental material also provides scripts and commands for reproducing the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized code is included in the supplemental material. It contains all the experiments reported in the paper. We only used the popular benchmark datasets, which are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup for reproducing our results is detailed in appendices. We submitted our source codes through OpenReview.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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 are repeated over multiple random seeds. The toy plot and result tables summarize repeated runs with variability information, while the tabular scatter plots display individual runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the hardware used for the experiments, including the CPU and GPU resources. The accompanying code package documents the software environment and package versions needed for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available benchmark datasets and focuses general machine learning methods. We do not release high-risk models or personally identifying information, and the study is conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive impacts, such as reducing the need for clean labels when evaluating and improving calibration, and potential risks in Appendix M

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new high-risk generative models, scraped datasets, or deployed systems requiring special safeguards. The experiments use existing public benchmark datasets and standard model architectures for calibration evaluation.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: [Yes]

Justification: The paper cites the original sources for all datasets and model assets used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The supplemental material includes an anonymized code package with a README describing the repository structure, dependencies, scripts, and commands needed to reproduce the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: We do not utilize such services, so the concerns raised in the question are not applicable to us.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve new crowdsourcing or new human-subject experiments requiring IRB approval. The empirical evaluation uses existing public benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as an important, original, or non-standard component of the core methodology. They were used only for writing, editing, formatting, and basic code-assistance purposes, which do not require declaration under the NeurIPS guidelines.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.