

Privacy Cost as Equity Input: A Group Fairness Criterion for Differentially Private Machine Learning

Rakshit Naidu

Georgia Institute of Technology
DATAmundi.ai
rakshitnaidu@gatech.edu

Abstract

Differential privacy (DP) is increasingly deployed to limit membership inference risk in machine-learning systems. Prior work has shown that DP-SGD can widen accuracy disparities across demographic groups, but this framing treats fairness as a purely outcome-side concern. We argue that privacy cost (denoted as the information leakage borne by each group) is itself a form of harm: a group that involuntarily bears greater privacy exposure is entitled to proportionally greater benefit from the system. From this principle we derive the Privacy-Cost Equity Ratio (PCER), a group fairness metric defined as a group’s positive prediction rate normalized by its per-group overfitting gap. By a standard membership inference bound, this overfitting gap upper-bounds each group’s advantage under the canonical threshold attack, making PCER a conservative measure of benefit relative to exposure. PCER requires only per-group train and test accuracy (no shadow models required) making it a practical post-hoc audit tool. We evaluate PCER alongside standard fairness metrics across six benchmark-attribute combinations spanning tabular and NLP domains, under DP-SGD at a range of privacy budgets. The results reveal patterns that outcome-based metrics miss. On Adult Income (race), PCER inverts the fairness ranking relative to standard metrics at the default floor constant: the privacy budget deemed most equitable by outcome gaps is not the most equitable by exposure-adjusted ones, though this specific ranking is itself sensitive to the floor constant. On COMPAS, PCER uncovers a persistently lower benefit-to-cost ratio for the protected group that demographic parity gap masks entirely. Sensitivity analysis further shows that very strong privacy guarantees collapse both groups’ overfitting to a numerical floor, rendering exposure-based audits uninformative in that regime. Taken together, these findings argue that fairness audits of privacy-preserving systems must account for who bears the cost of protection, as well as who benefits from its outcomes.

Extended version — <https://arxiv.org/abs/2607.16620>

1 Introduction

As machine-learning systems increasingly inform high-stakes decisions (creditworthiness, criminal recidivism prediction, income classification), practitioners face mounting pressure

to ensure these systems are simultaneously *fair* and *privacy-preserving*. Fairness is typically audited through outcome metrics such as demographic parity gap (DP-Gap) or equalized odds (EO-Gap); privacy is enforced through differential privacy (DP) (Dwork et al. 2006), with DP-SGD (Abadi et al. 2016) the dominant training-time mechanism against membership inference attacks (Shokri et al. 2017). The implicit assumption is that satisfying each demand independently suffices: a model passing a fairness audit *and* a certified DP budget is an acceptable deployment. We argue this is incomplete: what matters is whether the *distribution of privacy costs* across groups is itself equitable — a question neither audit asks in isolation.

A growing literature documents that DP-SGD hurts minority-group accuracy more than majority-group accuracy (Bagdasaryan, Poursaeed, and Shmatikov 2019; Uniyal et al. 2021; Tran, Fioretto, and Van Hentenryck 2021), typically framed as a fairness problem because *outcomes* worsen unevenly. But this framing treats privacy as a design constraint rather than a cost borne by individuals: when a membership inference attack succeeds on a training-set member, that person pays an *involuntary* privacy cost they did not consent to, and if some groups are systematically more exposed, they bear a larger share of it. Kulynych et al. (2022) confirm empirically that MIA success rates vary systematically across groups under DP-SGD, so privacy exposure is measurably non-uniform, despite the uniform guarantee DP’s ϵ might suggest.

Proportionality-based fairness criteria (Binns 2018) provide one normative lens to make this precise, among several we contrast in Section 3.2 (including equality-of-treatment and Rawlsian maximin alternatives): a system is equitable when the ratio of prediction benefit to involuntary privacy cost is equal across groups, so that a group contributing more privacy exposure receives proportionally greater benefit as compensation. This is a substantive ethical commitment, not a mathematical necessity, and we adopt it because it is well suited to *retroactive* audits of already-incurred, involuntary harm; readers who favor a different normative starting point can still use the overfitting-gap measurements PCER produces to instantiate their own criterion. Under this lens, unequal benefit-to-cost ratios are invisible to standard fairness metrics, which check only whether outcome rates are equal, not whether they are proportional to involuntary exposure.

The practical consequences of this audit gap are significant. Deployers certifying both a fairness audit (DP-Gap) and a privacy audit (ϵ) may be unaware that one demographic group bears a systematically higher fraction of realized privacy cost, invisible to either metric alone: DP-Gap ignores exposure asymmetry, and equalized odds treats privacy as a separate concern entirely. A budget ranking optimal under demographic parity can simultaneously impose a disproportionate privacy burden on a historically disadvantaged group. As regulation increasingly demands auditable fairness evidence (the EU AI Act mandates bias monitoring for high-risk systems (European Parliament and Council of the European Union 2024)), the absence of a privacy-cost equity metric leaves a concrete, actionable gap practitioners cannot currently fill.

Figure 1 illustrates why these considerations matter in practice. On Adult Income (race, panel a), choosing $\epsilon \in \{5, 10\}$ minimizes DP-Gap, the standard fairness criterion, yet at the default floor ($\eta = 10^{-3}$) these are *not* the most equitable budgets by the proportionality-based metric we introduce, which is minimized at no-DP. The two metrics recommend different deployment decisions at this floor; Whites enjoy a significantly positive benefit-to-cost ratio at every budget (CIs exclude zero throughout), but *which* budget PCER ranks as most equitable is itself floor-sensitive (see the extended version). On COMPAS (race, panel b), DP-Gap is nearly flat across most budgets, suggesting DP provides little additional fairness benefit at any ϵ . But the proportionality metric is persistently and significantly negative (95% bootstrap CI excludes zero at every budget, with stable, floor-robust estimates at $\epsilon \geq 0.5$; 50 seeds), revealing that African-Americans receive a lower benefit-to-cost ratio than the comparison group, a pattern DP-Gap misses entirely.

In this paper, we close this gap by introducing the *Privacy-Cost Equity Ratio* (PCER), a group fairness metric operationalizing harm-proportionality via Adams’ equity-ratio structure (Adams 1963): the ratio of a group’s positive prediction rate to its per-group overfitting gap $\delta_g = \max(0, \text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}})$, which by the Yeom et al. (Yeom et al. 2018) bound upper-bounds per-group membership inference advantage, making it a theoretically grounded, shadow-model-free privacy-cost proxy. A system is PCER-equitable when this ratio is equal across groups. We evaluate PCER against standard metrics on four benchmark datasets (six benchmark-attribute combinations) under DP-SGD, validate the overfitting-gap proxy against a direct threshold membership-inference attack, provide a floor-constant sensitivity analysis, and package the procedure as a deployable post-hoc audit (Algorithm 1).

Our contributions are as follows:

1. We formally define *Privacy-Cost Equity Ratio* (PCER) using the per-group overfitting gap as the privacy-cost proxy, grounded in the Yeom et al. bound connecting overfitting to MIA advantage, and show PCER is normatively distinct from demographic parity and equalized odds (Section 3).
2. We establish three theoretical results that ground PCER’s uniqueness and structural properties (Section 3): an *axiomatic characterization* showing R_g/δ_g is the unique

benefit-to-cost mapping satisfying linearity in benefit and inverse linearity in cost (proof in the extended version); a *floor degeneracy theorem* (Proposition 2) proving that PCER reduces to scaled DP-Gap precisely when both groups’ overfitting gaps fall below the floor; and a *PCER–DP-Gap incompatibility theorem* (Proposition 3) showing the two fairness criteria cannot hold simultaneously unless per-group overfitting gaps are equal, a condition DP-SGD does not guarantee.

3. We demonstrate empirically on Adult Income, COMPAS, ACS Income 2018, and Bias in Bios (NLP domain; 50 seeds each; six benchmark-attribute combinations) that PCER diverges from DP-Gap and EO-Gap at moderate DP budgets ($\epsilon \in \{1, 5, 10\}$), with the *sign* of these disagreements robust across floor choices and with bootstrap CIs excluding zero (Section 4).
4. We conduct a floor-constant sensitivity analysis ($\eta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$) that shows PCER findings at strong DP ($\epsilon = 0.1$) are floor-dominated artifacts, while findings at $\epsilon \geq 1$ are robust to floor choice (Section 5).
5. We empirically validate the overfitting-gap proxy against a direct, per-group threshold membership-inference attack across all six benchmark-attribute combinations and DP budgets, finding strong directional agreement (Section 5.5).

Practical significance. Three properties make PCER a viable addition to existing audit pipelines. First, it requires only per-group train/test accuracy (no shadow models), so its marginal cost over an existing DP-SGD evaluation pipeline is negligible compared to LiRA (Carlini et al. 2022)-style approaches requiring hundreds of shadow-model trainings. Second, PCER is, to our knowledge, the only audit tool that surfaces **double disadvantage** (a group bearing both higher privacy exposure *and* worse predictive outcomes) when it occurs. DP-Gap cannot surface this because it has no denominator sensitive to differential memorization. Third, Proposition 3 gives auditors a falsifiable claim: any system asserting compliance with both a DP-Gap and a PCER audit is implicitly claiming equal per-group overfitting, an assertion Algorithm 1 can confirm or refute with no additional data collection.

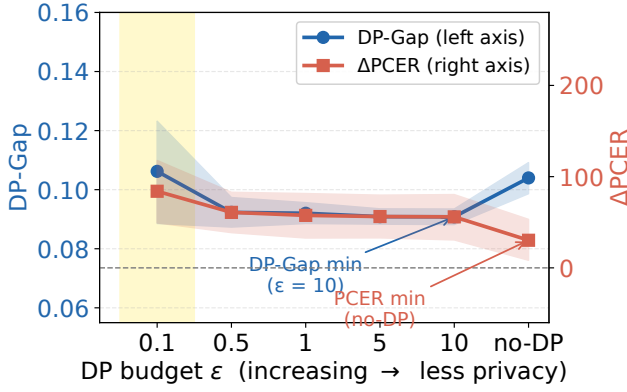
2 Background and Related Work

2.1 Adams’ Equity Theory and Proportionality-Based Fairness

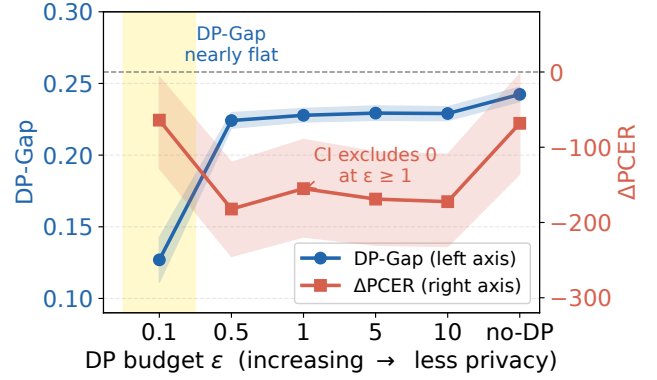
Adams (Adams 1963) proposed that individuals evaluate fairness by comparing their own outcome-to-input ratio to a referent’s:

$$\frac{\text{outcome}_i}{\text{input}_i} = \frac{\text{outcome}_j}{\text{input}_j}.$$

Inequitable ratios produce tension motivating corrective action. The theory was originally framed around voluntary inputs (effort, skill), but Binns (2018) observes that Adams’ proportionality structure motivates fairness criteria in which the “input” need not be voluntary: any systematic cost borne by a party can ground a proportionality claim. Speicher et al.



(a) Adult Income (race), 50 seeds, 95% CI. DP-Gap (blue, left axis) is minimized at $\varepsilon \in \{5, 10\}$. Standard fairness criteria select these as the most equitable budgets. At the default floor ($\eta = 10^{-3}$), Δ_{PCER} (red, right axis) is minimized at no-DP, but which budget PCER prefers is floor-sensitive (see the extended version); the two metrics nonetheless produce conflicting recommendations at this floor.



(b) COMPAS (race), 50 seeds, 95% CI. DP-Gap (blue) is nearly flat across most DP budgets. Δ_{PCER} (red) is persistently negative with CI excluding zero at every budget and stable at $\varepsilon \geq 0.5$ (the shaded $\varepsilon = 0.1$ estimate is floor-dominated): African-Americans consistently receive a lower benefit-to-cost ratio than the comparison group (a higher harmful-prediction rate relative to overfitting exposure), invisible to DP-Gap.

Figure 1: Two findings that standard fairness metrics cannot detect. Each panel shows DP-Gap and Δ_{PCER} (with 95% bootstrap CI) as a function of DP budget ε , with no-DP ($\varepsilon = \infty$) on the right. Yellow shading marks the floor-dominated regime ($\varepsilon = 0.1$).

(2018) similarly propose a family of metrics unified by an “individual goodness” principle rewarding outcome-merit alignment. These accounts identify the proportionality requirement but stop short of specifying which cost should enter the denominator in an ML context; to our knowledge, PCER is the first to instantiate Adams’ broader reading with involuntary privacy exposure as that cost.

2.2 Differential Privacy and DP-SGD

A randomized mechanism $\mathcal{M}$ satisfies (ε, δ) -differential privacy (Dwork et al. 2006) if for any neighboring datasets D, D' and output set S :

$$\Pr[\mathcal{M}(D) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in S] + \delta.$$

DP-SGD (Abadi et al. 2016) achieves DP for neural-network training by clipping per-sample gradients to norm C and adding Gaussian noise $\mathcal{N}(0, \sigma^2 C^2 I)$ at each step, with a privacy accountant tracking the cumulative (ε, δ) budget. A growing literature documents that DP-SGD does not affect all groups equally: DP noise disproportionately degrades accuracy for underrepresented groups (Bagdasaryan, Poursead, and Shmatikov 2019), motivating fair DP training that constrains equalized-odds violations within a budget (Tran, Fioretto, and Van Hentenryck 2021), and characterizations of the accuracy–fairness–privacy trilemma (Jagielski et al. 2019). These works treat fairness as an outcome-rate property; PCER instead asks whether DP budgets are equitable with respect to the *costs* DP imposes on different groups.

2.3 Membership Inference and Overfitting

A membership inference attack (MIA) (Shokri et al. 2017) determines whether a target record was in the training set. The canonical threshold attack (Yeom et al. 2018) scores by

per-sample loss (low loss indicates memorization) and its advantage is bounded above by the generalization gap, connecting overfitting directly to privacy leakage. We apply this bound per group, yielding an auditable, shadow-model-free bound on group-level MIA advantage. Kulynych et al. (2022) provide complementary empirical evidence that DP mechanisms expose demographic groups to different membership inference risk under the same nominal ε , persisting across DP budgets, motivating a criterion that normalizes benefit by per-group privacy cost. We adopt the overfitting gap for this; stronger attacks such as LiRA (Carlini et al. 2022) can tighten the bound but require expensive shadow-model training not necessary for audit deployment.

2.4 Group Fairness Criteria and Their Limits

The two most widely used group fairness metrics are:

- **Demographic Parity Gap (DP-Gap):** $|\Pr[\hat{Y} = 1 \mid A = 0] - \Pr[\hat{Y} = 1 \mid A = 1]|$. Requires equal positive prediction rates (Dwork et al. 2012).
- **Equalized Odds Gap (EO-Gap):** $\max(\text{TPR gap}, \text{FPR gap})$. Requires equal true/false positive rates (Hardt, Price, and Srebro 2016).

Both metrics evaluate *outcome disparities* without reference to costs borne by each group; EO-Gap additionally conditions on the true label, but neither captures whether outcome rates are proportional to involuntary privacy exposure. Chang and Shokri find a tension between enforced outcome-rate parity and membership inference risk (Chang and Shokri 2021); Cummings et al. study the formal compatibility of individual privacy and group fairness (Cummings et al. 2019). Our work is complementary: rather than studying how fairness constraints affect privacy, we study how DP budget choices

affect a *combined* criterion sensitive to both. PCER operates at the group level and is orthogonal to the individual fairness of (Dwork et al. 2012).

3 Methodology

3.1 Formal Definitions

Let $f_\theta : \mathcal{X} \rightarrow [0, 1]$ be a probabilistic classifier trained on D_{train} , and $A \in \{0, 1\}$ a binary protected attribute.

Definition 1 (Group Positive Rate). *The positive prediction rate for group g on test set D_{test} is*

$$R_g = \mathbb{E}[\mathbf{1}[f_\theta(X) \geq 0.5] \mid A = g, (X, Y) \sim D_{\text{test}}].$$

Remark 1 (Positive rate as benefit proxy). R_g approximates benefit delivery in beneficial-prediction domains and harm received in harmful-prediction domains (e.g. criminal recidivism); it reflects both base-rate differences and model-induced disparities (Section 6).

Definition 2 (Per-Group Overfitting Gap). Let $\text{acc}_g^{\text{tr}} = \mathbb{E}[\mathbf{1}[\mathbf{1}[f_\theta(X) \geq 0.5] = Y] \mid A = g, (X, Y) \sim D_{\text{train}}]$ and $\text{acc}_g^{\text{te}} = \mathbb{E}[\mathbf{1}[\mathbf{1}[f_\theta(X) \geq 0.5] = Y] \mid A = g, (X, Y) \sim D_{\text{test}}]$. The per-group overfitting gap is

$$\delta_g = \max(0, \text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}}).$$

Proposition 1 (Yeom bound on per-group MIA advantage). *Fix the specific threshold-based membership inference attacker of Yeom et al. (Yeom et al. 2018), which decides membership for record i in group g by thresholding its loss ℓ_i at (a group-specific estimate of) the expected training loss. Let $\hat{\mu}_g$ denote this attacker’s accuracy-based advantage on group g , i.e.*

$$\hat{\mu}_g = \Pr[\text{attacker_is_correct}] - \frac{1}{2},$$

evaluated on a balanced member/non-member set restricted to group g . Then

$$\hat{\mu}_g \leq \delta_g.$$

Proof. Yeom et al. (Yeom et al. 2018) prove that their threshold attacker achieves $\Pr[\text{correct}] \leq \frac{1}{2} + \frac{1}{2}(L_g^{\text{te}} - L_g^{\text{tr}})$, where L_g is the expected 0-1 loss (error rate) on group g . Since $\hat{\mu}_g = \Pr[\text{correct}] - \frac{1}{2}$ and $L_g^{\text{te}} - L_g^{\text{tr}} = \text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}}$, substituting gives the tighter bound $\hat{\mu}_g \leq \frac{1}{2}(\text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}}) \leq \frac{1}{2}\delta_g$, which is twice as tight as the inequality stated above. We report the coarser $\hat{\mu}_g \leq \delta_g$ throughout because δ_g — without the factor of $\frac{1}{2}$ — is the quantity used directly as PCER’s denominator (Definition 3); the stated proposition is the valid, looser guarantee that choice of denominator actually provides. $\max(\cdot, 0)$ clips at zero for non-overfit groups. This bound is specific to the fixed-threshold adversary evaluated at accuracy, not to threshold-independent summaries such as ROC-AUC, and stronger attacks (e.g. LiRA (Carlini et al. 2022)) may achieve higher advantage not captured here. Section 5.5 reports AUC-based attack success as a complementary empirical diagnostic without claiming it is bounded by δ_g in this same formal sense. $\square$

Definition 3 (Privacy-Cost Equity Ratio). *For group g with positive rate R_g and overfitting gap δ_g , the Privacy-Cost Equity Ratio is*

$$\text{PCER}_g = \frac{R_g}{\max(\delta_g, \eta)},$$

where $\eta > 0$ is a small positive floor constant (default $\eta = 10^{-3}$; see the extended version for the full sensitivity analysis).

Remark 2 (Conservative lower bound on true per-group PCER). *By Proposition 1, $\hat{\mu}_g \leq \delta_g$, so $\max(\hat{\mu}_g, \eta) \leq \max(\delta_g, \eta)$, giving*

$$\text{PCER}_g = \frac{R_g}{\max(\delta_g, \eta)} \leq \frac{R_g}{\max(\hat{\mu}_g, \eta)}.$$

Using δ_g in the denominator overestimates group g ’s privacy cost and therefore underestimates its true PCER value. This per-group bound does not automatically transfer to the signed disparity $\Delta_{\text{PCER}} = \text{PCER}_0 - \text{PCER}_1$: because the two groups’ PCER values are underestimated by different, uncontrolled amounts, $|\Delta_{\text{PCER}}^\delta| \leq |\Delta_{\text{PCER}}^{\text{MIA}}|$ is not a corollary of this remark, only a pattern we test empirically (Section 5.5), where it holds in most but not all configurations.

PCER’s ratio structure is forced: it is the unique benefit-to-cost mapping satisfying linearity in benefit, inverse linearity in cost, and a scale-normalization convention (the extended version gives the formal axiomatization and proof), ruling out alternatives such as an unnormalized difference $R_g - \delta_g$ or a super-linear cost penalty R_g/δ_g^2 .

Proposition 2 (Floor Degeneracy). *Let $\eta > 0$ be fixed. If $\delta_0, \delta_1 < \eta$, then*

$$\Delta_{\text{PCER}} = \frac{R_0 - R_1}{\eta} = \frac{\text{DP-Gap (signed)}}{\eta}.$$

PCER provides no diagnostic information beyond a scaled version of DP-Gap in this regime.

Proof. When $\delta_g < \eta$, $\max(\delta_g, \eta) = \eta$ for both $g \in \{0, 1\}$. Therefore $\Delta_{\text{PCER}} = R_0/\eta - R_1/\eta = (R_0 - R_1)/\eta$. $\square$

Remark 3 (Floor-dominated regime under strong DP). *Under DP-SGD, sufficiently strong noise ($\varepsilon \ll 1$) drives the model toward near-random predictions, collapsing per-group overfitting toward zero. Proposition 2 formalizes the consequence: once $\delta_g < \eta$ for all g , a PCER audit is equivalent to a scaled DP-Gap audit and should be reported as such. The floor sensitivity analysis (detailed in the extended version) operationalizes this check: if $|\Delta_{\text{PCER}}|$ scales proportionally with $1/\eta$ across orders of magnitude, the finding is floor-dominated. All findings at $\varepsilon = 0.1$ in our experiments fall in this regime.*

Definition 4 (PCER Disparity). *The signed PCER disparity for groups $\{0, 1\}$ is*

$$\Delta_{\text{PCER}} = \text{PCER}_0 - \text{PCER}_1.$$

A system is PCER-equitable when $\Delta_{\text{PCER}} = 0$.

Algorithm 1 requires only a held-out test set and evaluation of the already-trained model on the training set. Neither shadow models nor auxiliary attack infrastructure are required, which makes PCER deployable as a post-hoc audit step in any existing DP-SGD pipeline.

Algorithm 1: PCER Audit for DP Budget Selection

Part I — Trainer *trains models; releases per-group statistics*

Require: Data D , protected attribute A , candidate budgets $\mathcal{E}$

- 1: **for** each $\varepsilon \in \mathcal{E}$ **do**
- 2: Train f^ε on D_{train} with DP-SGD at budget ε
- 3: **for** $g \in \{0, 1\}$ **do**
- 4: **Release** $R_g^\varepsilon \leftarrow \mathbb{E}[\mathbf{1}[f^\varepsilon(X) \geq 0.5] \mid A=g, (X, Y) \sim D_{\text{test}}]$
- 5: **Release** $\text{acc}_g^{\text{tr}}(f^\varepsilon)$ and $\text{acc}_g^{\text{te}}(f^\varepsilon)$
- 6: **end for**
- 7: **end for**

Part II — Auditor *uses only released statistics; no model or data access required*

Require: Released $\{R_g^\varepsilon, \text{acc}_g^{\text{tr}}, \text{acc}_g^{\text{te}}\}_{g,\varepsilon}$, floor grid $\mathcal{H} = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$

Ensure: Per-floor equitable budget $\varepsilon^*(\eta')$; robustness verdict

- 8: **for** each $\varepsilon, g \in \{0, 1\}$ **do**
- 9: $\delta_g^\varepsilon \leftarrow \max(0, \text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}})$
- 10: **end for**
- 11: **for** each $\eta' \in \mathcal{H}$ **do**
- 12: **for** each ε **do**
- 13: $\Delta^\varepsilon(\eta') \leftarrow \left| \frac{R_0^\varepsilon}{\max(\delta_0^\varepsilon, \eta')} - \frac{R_1^\varepsilon}{\max(\delta_1^\varepsilon, \eta')} \right|$
- 14: **end for**
- 15: $\varepsilon^*(\eta') \leftarrow \arg \min_\varepsilon \Delta^\varepsilon(\eta')$
- 16: **end for**
- 17: **if** $|\{\varepsilon^*(\eta') : \eta' \in \mathcal{H}\}| = 1$ **then**
- 18: **Report** ε^* as the PCER-equitable budget $\triangleright$
- 19: **else**
- 20: **Flag:** budget ranking is floor-sensitive; inspect $\Delta^\varepsilon(\eta')$ table
- 21: **end if**

3.2 Normative Justification

PCER operationalizes a harm-proportionality equity principle using Adams’ output-to-input ratio framework:

- **Output (benefit):** positive prediction rate R_g , the benefit (or harm, in domains like criminal recidivism) the model delivers to group g .
- **Input (involuntary cost):** overfitting gap δ_g , the degree to which the model has memorized group g ’s training data beyond what is needed for generalization. By Proposition 1, this directly bounds the privacy leakage group g is exposed to.

The key distinction from Adams’ original formulation is that the “input” is involuntary: data subjects cannot choose how much of their data the training process memorizes. Memorization is the mechanism of privacy leakage (Yeom et al. 2018); groups whose data is memorized more are at greater risk of membership inference. The proportionality norm we invoke is therefore one of *compensatory fairness* (Binns 2018): when

a joint enterprise imposes unequal involuntary costs on different groups, those who bear more cost are owed proportionally more benefit as compensation.

Formally, the principle requires $\text{PCER}_0 = \text{PCER}_1$, i.e. $R_0/\delta_0 = R_1/\delta_1$. A PCER violation ($\Delta_{\text{PCER}} \neq 0$) means the groups receive unequal benefit-to-memorization-cost ratios. The group with the lower ratio is doubly disadvantaged, bearing higher privacy risk for less outcome benefit.

Alternatives we considered. PCER’s compensatory-proportionality norm is one candidate among several; we do not claim it is the unique correct standard. *Equality of treatment*, the standard that DP-Gap and EO-Gap operationalize (Section 3.3), holds a system fair once outcome rates are equalized regardless of how unevenly privacy cost fell on each group; it is silent on exposure asymmetry, which a compensatory view treats as morally relevant. A *Rawlsian maximin* criterion (maximize the worst-off group’s welfare) is designed for *ex ante* contract design under a veil of ignorance (Binns 2018); our setting is retroactive, so compensatory restitution fits better than a prospective redistribution rule. *Capability-based* (Nussbaum 2011) and *distributive-justice* accounts would instead ask whether groups can achieve functionings they value, or prioritize equalising outcomes over compensating differential costs. Practitioners who reject the premise that involuntary cost grounds compensation will reasonably prefer DP-Gap alone and should read PCER as diagnostic rather than a corrected successor metric; its contribution is to make the exposure dimension *measurable*, whichever normative standard is ultimately applied.

Why compensation rather than equalization? If unequal memorization is the injustice, the remedy should be to equalize δ_g across groups directly. But DP-SGD provides only a mechanism-level guarantee on the aggregate budget ε , with no per-group memorization control: the overfitting disparity arises from dataset statistics (group size, feature density, label separability) that cannot be corrected without altering the training distribution. Where equalization is infeasible, compensatory proportionality is the appropriate second-best norm, and PCER makes it auditable. Like demographic parity, PCER is a group criterion aggregating over individuals within each demographic group; it does not imply individual fairness in the sense of (Dwork et al. 2012).

3.3 Relationship to Demographic Parity Gap

DP-Gap measures $|R_0 - R_1|$; Δ_{PCER} measures $R_0/\delta_0 - R_1/\delta_1$. The two coincide only when $\delta_0 = \delta_1$. Under DP-SGD, the overfitting gap is not group-invariant, i.e. minority groups with fewer training samples can exhibit either higher or lower δ_g than majority groups, depending on the noise schedule and group-specific signal strength, so the metrics can diverge. When both groups’ gaps are at floor η , Δ_{PCER} reduces to DP-Gap/ η (Proposition 2).

Away from the floor, PCER and DP-Gap can disagree in *direction* as well as magnitude, precisely when the group with the higher outcome rate bears a proportionally higher memorization cost (proof in the extended version); this requires genuinely unequal overfitting gaps, so at the floor the two metrics necessarily agree in direction (Proposition 2).

Proposition 3 (PCER–DP-Gap Incompatibility). *Assume $R_0, R_1 > 0$ and $\delta_0, \delta_1 \geq \eta$, and suppose DP-Gap-equity holds ($R_0 = R_1$). Then, conditional on DP-Gap-equity, PCER-equity ($\Delta_{\text{PCER}} = 0$) holds if and only if $\delta_0 = \delta_1$.*

Proof. ($\Rightarrow$) Suppose DP-Gap-equity holds and PCER-equity also holds. DP-Gap-equity gives $R_0 = R_1 = R > 0$. Since $\delta_0, \delta_1 \geq \eta$ by assumption, $\max(\delta_g, \eta) = \delta_g$ for $g \in \{0, 1\}$, so $\text{PCER}_g = R/\delta_g$. Substituting into $\Delta_{\text{PCER}} = 0$ gives $R/\delta_0 = R/\delta_1$. Because $R > 0$, the map $x \mapsto R/x$ is injective on $x > 0$, so this equality of images forces equality of the arguments: $\delta_0 = \delta_1$.

($\Leftarrow$) Suppose DP-Gap-equity holds ($R_0 = R_1 = R$) and $\delta_0 = \delta_1 =: \delta$. Then $\text{PCER}_0 = R/\max(\delta, \eta) = \text{PCER}_1$, so $\Delta_{\text{PCER}} = 0$ trivially. Combining both directions establishes the claimed biconditional *conditional on DP-Gap-equity already holding*; without that premise, $\delta_0 = \delta_1$ alone does not force $R_0 = R_1$, so the two equities are independent facts about R and δ respectively, linked only through this conditional statement. $\square$

Since DP-SGD gives no mechanism-level control over $\delta_0 \neq \delta_1$ (Section 3.2), the two audits are in general jointly unsatisfiable; Section 6 returns to what this means for auditors.

4 Experiments

4.1 Datasets

UCI Adult Income. UCI Adult Income ($n = 46,033$ after preprocessing) (Becker and Kohavi 1996) with protected attributes *sex* (Male/Female) and *race* (White/Non-White); binary outcome *income > \$50K/year* (beneficial). Features: age, education, occupation, hours worked, capital gain/loss; protected attributes excluded from features.

COMPAS. ProPublica COMPAS recidivism ($n = 6,172$) (Angwin et al. 2016) with race as protected attribute (African-American vs. Other); outcome is two-year recidivism, a *harmful* prediction. Features: age, prior count, charge degree, juvenile offense counts.

ACS Income 2018. ACS 1-Year PUMS 2018 via folktables (Ding et al. 2021), states CA, TX, NY (combined $n \approx 843,000$, subsampled to $n = 50,000$); task mirrors Adult Income (predict *income > \$50K/year*). Two protected attributes evaluated independently: *race* (RACIP= 1 $\rightarrow$ White, else Non-White) and *sex* (SEX= 1 $\rightarrow$ Male, SEX= 2 $\rightarrow$ Female), excluded from features.

Bias in Bios. Bias in Bios (De-Arteaga et al. 2019) ($n = 37,900$ after filtering and class-balancing the professor/nurse task): online biographies with occupation labels and binary gender, predicting *professor* ($y = 1$, beneficial) vs. *nurse* ($y = 0$); gender (Male= 0, Female= 1) is the protected attribute. We use the `hard_text` field (name removed) to prevent name-to-gender shortcuts (Romanov et al. 2019), encoded offline via TF-IDF (unigrams+bigrams, 15,000 vocabulary, sublinear TF) and Truncated SVD to 200 components, fed to the same MLP architecture as the tabular experiments.

This tests whether PCER findings replicate in the NLP domain, where the beneficial outcome is access to a high-status occupational label rather than a financial one.

4.2 Experimental Setup

We train a three-layer MLP (64–32–1 units, ReLU) using DP-SGD (Abadi et al. 2016) via Opacus 1.5 (Yousefpour et al. 2021), sweeping $\varepsilon \in \{0.1, 0.5, 1.0, 5.0, 10.0\}$ with a non-private (no-DP) baseline. All runs use $\delta = 10^{-5}$, max gradient norm $C = 1.0$, batch size 256, SGD with learning rate 0.05 and momentum 0.9, and 25 training epochs. Features are standardized. We use a 70/30 stratified train/test split. All datasets use 50 independent seeds.

Privacy cost is measured via the per-group overfitting gap $\delta_g = \max(0, \text{acc}_g^{\text{tr}} - \text{acc}_g^{\text{te}})$, computed by evaluating the trained model on the train and test splits per group. No shadow models are required. Floor constant $\eta = 10^{-3}$ unless otherwise stated; sensitivity to η is analysed in the extended version.

Direct threshold-MIA validation protocol. To validate δ_g against an actual attack rather than only its theoretical bound, we additionally run the classical per-group threshold membership-inference attack (Yeom et al. 2018) on every trained model above (same seeds, ε settings, and six benchmark-attribute combinations; no additional training required). For each group g we score every train/test example by negative loss $-\ell_i$, pool member/non-member scores, and report AUC_g (area under the ROC curve; 0.5 = chance) as a threshold-free summary of attack success, kept deliberately separate from the accuracy-based $\hat{\mu}_g$ that Proposition 1 bounds (we make no formal claim δ_g bounds AUC_g in the same sense). The empirical question: does the overfitting-gap proxy agree, directionally and in the PCER rankings it implies, with a real attacker’s success rate? Results appear in Section 5.5.

4.3 Evaluation Criteria

We assess divergence along three axes:

1. **Ranking disagreement.** Whether the two metrics order the ε settings differently in terms of which is “most fair.”
2. **Directional disagreement.** Whether the metrics disagree on *which group* is disadvantaged at any ε setting (see the extended version).
3. **Floor sensitivity.** Whether findings at $\varepsilon = 0.1$ persist or vanish as η varies (extended version).

Pearson correlations between Δ_{PCER} and $\Delta_{\text{DP-Gap}}$ are computed across ε settings (5 data points for the full sweep, 4 for $\varepsilon \geq 1$ only). These correlations have very limited statistical power and are reported as descriptive summaries rather than inferential results.

5 Results

Adult Income (sex and race) results are reported in the extended version: both attributes show robust positive Δ_{PCER} (CIs excluding zero at all $\varepsilon \geq 1$), and race additionally shows a ranking reversal at the default floor (DP-Gap prefers $\varepsilon \in \{5, 10\}$, PCER prefers no-DP); the positive sign of

Δ_{PCER} persists at $\eta = 10^{-2}$, though which budget PCER ranks as most equitable is itself floor-sensitive.

5.1 COMPAS: Race

Table 1 reports DP-Gap, EO-Gap, per-group overfitting gaps, and Δ_{PCER} across the DP budget sweep on COMPAS (Angwin et al. 2016).

Floor-dominated regime at $\varepsilon = 0.1$. DP-Gap identifies $\varepsilon = 0.1$ as the most equitable budget (mean 0.127 ± 0.059). PCER’s mean-based CI here technically excludes zero (-64 ± 220 , CI $[-128, -5]$), yet only 28/50 seeds (56%) are negative, indistinguishable from a fair coin ($p = 0.48$). The two tests disagree because $\Delta_{\text{PCER}} = R/\delta$ has a noisy, near-zero denominator at this budget ($\delta_g \approx .011$ – $.013$, std comparable to the mean), so a few outlier seeds pull the mean below zero without a consistent per-seed sign; we treat mean-CI and sign test as complementary, not interchangeable. Raising η makes the sign more consistent (36/50, then 48/50 negative), but per Proposition 2 only because Δ_{PCER} degenerates into a rescaled DP-Gap once $\delta_g \ll \eta$, not from a genuine effect: no η makes this budget informative (Remark 3). At $\varepsilon \geq 0.5$ below, mean-CI and sign agree, so we report the mean-CI alone.

Robust findings at $\varepsilon \geq 0.5$. PCER at $\varepsilon \geq 0.5$ is stably and significantly negative, floor-robust, and in close agreement across budgets: African-Americans consistently bear a higher harm-to-cost ratio. The $\varepsilon = 0.5$ estimate is most extreme because the African-American overfitting gap is smaller there than at $\varepsilon = 1$ (.0060 vs. .0068), shrinking that group’s denominator while outcome rates stay comparable. EO-Gap tracks DP-Gap closely and is nearly flat across $\varepsilon \geq 1$ (0.239–0.242), while PCER remains stable at approximately -160 . Per-group exposure is comparable in this regime (δ_{Ot} and δ_{AA} both $\approx .006$, with neither group consistently higher), so the disparity is carried by African-Americans’ higher harmful-prediction rate relative to that shared exposure, and DP noise does not reduce it (most robust at $\varepsilon = 5$: -169 , CI $[-230, -106]$; floor-robust -19.9 ± 11.7 at $\eta = 10^{-2}$).

5.2 Floor-Constant Sensitivity: Summary

Full floor-constant sensitivity results are reported in the extended version. The key pattern: at $\varepsilon = 0.1$, both magnitude and direction of Δ_{PCER} are floor-sensitive, as predicted by Proposition 2 ($10\times$ change in η scales $|\Delta_{\text{PCER}}|$ by $\sim 10\times$; positive-direction seeds shift 22/50 $\rightarrow$ 14/50 $\rightarrow$ 2/50). At $\varepsilon \geq 1$, direction is stable across all $\eta \in \{10^{-4}, \dots, 10^{-1}\}$, confirming a genuine differential rather than a floor artifact.

5.3 ACS Income 2018

Table 2 reports the race and sex attributes on ACS Income 2018 (Ding et al. 2021).

The ACS Income 2018 dataset (Table 2) provides a larger-scale replication ($n = 50,000$) of the race and sex patterns observed on Adult.

ACS race. PCER has large standard deviation (all CIs cross zero) at all ε , reflecting small, similar overfitting gaps across groups ($\delta_{\text{W}} \approx 0.004$ – 0.007 , $\delta_{\text{NW}} \approx 0.003$ – 0.006); direction

alternates sign across ε (-13 at no-DP, $+12$ at $\varepsilon = 0.1$, -14 at $\varepsilon = 1$). At $\eta = 10^{-2}$ direction is consistently positive ($+5.4$ to $+5.7$) but no estimate excludes zero at $\eta = 10^{-3}$. *No robust PCER divergence from DP-Gap is established for ACS race.*

ACS sex. DP-Gap on ACS is much smaller than on Adult sex (0.079–0.097 vs. 0.184–0.188). PCER is positive at all ε (sign consistent with Adult sex) but CIs cross zero throughout; at $\eta = 10^{-2}$ direction is positive and tighter ($+8.2$ to $+9.7$), consistent with the Adult sex pattern but not statistically robust. *No robust PCER finding is observed for ACS sex at any ε .*

5.4 Bias in Bios: Gender

Table 3 reports results for the NLP-domain benchmark. Bias in Bios exhibits a structural gender asymmetry (males predicted professor at 86%, females at 33%), yielding a large DP-Gap (>0.5) across $\varepsilon \geq 1$ that reflects class imbalance inherited from the gender-skewed occupational distribution, not a model-induced artifact (De-Arteaga et al. 2019).

No-DP baseline: tightest CI in the study. At no-DP, $\Delta_{\text{PCER}} = +22.2 \pm 3.4$, CI $[+21.3, +23.2]$, the tightest interval across all six benchmarks, since both overfitting gaps are large and stable ($\delta_{\text{M}} = 0.029 \pm 0.003$, $\delta_{\text{F}} = 0.045 \pm 0.002$). Per-group PCER is $R_{\text{M}}/\delta_{\text{M}} \approx 29.3$ vs. $R_{\text{F}}/\delta_{\text{F}} \approx 7.5$: although females’ records are memorized more ($\delta_{\text{F}} > \delta_{\text{M}}$), they also receive a lower professor-prediction rate, so males receive more benefit per unit of privacy cost, an inequity present even absent DP (50/50 seeds positive).

Floor-dominated at $\varepsilon = 0.1$, robust positive at $\varepsilon \geq 0.5$. At $\varepsilon = 0.1$, both gaps collapse near zero and the model degrades to near-chance accuracy, yielding $\Delta_{\text{PCER}} = +65 \pm 316$ (CI $[-17, +152]$): floor-dominated and uninformative. At $\varepsilon \geq 0.5$, PCER is consistently positive ($+349$ down to $+209$, all CIs excluding zero; $\varepsilon = 0.5$ peaks from a smaller denominator than $\varepsilon = 1$), direction robust (43/50 to 50/50 seeds positive) and surviving at $\eta = 10^{-2}$ per the extended version. These findings replicate the Adult (sex) pattern in an NLP setting, including at no-DP, where the occupational disparity predates DP-SGD and is amplified as noise compresses the overfitting denominator.

5.5 Validation Against Direct Threshold-MIA Attacks

Proposition 1 and Remark 2 justify δ_g as a theoretical upper bound on attack advantage, but a bound is not a measurement. We therefore additionally run a direct per-group threshold membership-inference attack on every trained model (Section 4) and compare the PCER disparity it implies, $\Delta_{\text{PCER}}^{\text{MIA}}$ (substituting $\max(\text{AUC}_g - 0.5, \eta)$ for $\max(\delta_g, \eta)$), against $\Delta_{\text{PCER}}^\delta$; full results are reported in the extended version (split across two tables by dataset).

Direct attack success is weak; the proxy tracks it directionally. Pooling all six combinations and six DP budgets (72 group-level observations), the threshold attack’s raw advantage $\text{AUC}_g - 0.5$ has mean 0.0013 and maximum 0.0118,

ε	Acc	DP-Gap	EO-Gap	δ_{OI}	δ_{AA}	Δ_{PCER}
no-DP	.682±.008	.242±.018	.246±.030	.0074±.010	.0099±.012	−68±245 [−135, −2]
0.1	.590±.043	.127±.059	.134±.050	.0109±.011	.0127±.014	−64±220 [−128, −5]
0.5	.675±.007	.224±.020	.237±.029	.0063±.0084	.0060±.0097	−182±232 [−245, −120]
1.0	.678±.008	.228±.017	.239±.027	.0058±.008	.0068±.010	−155±240 [−219, −90]
5.0	.679±.008	.229±.018	.242±.030	.0065±.009	.0065±.011	−169±229 [−230, −106]
10.0	.679±.008	.229±.018	.242±.030	.0063±.009	.0061±.010	−172±231 [−234, −109]

Table 1: COMPAS, race attribute (Other=0, African-American=1). Mean $\pm$ std over **50** independent runs (seeds 0–49). δ_g : per-group overfitting gap; Δ_{PCER} : PCER disparity (95% bootstrap CI in brackets). Positive prediction = predicted recidivism (**harmful**). $\Delta_{PCER} < 0$ means African-Americans bear higher harm-to-cost ratio. **Bold**: minimum DP-Gap. The $\varepsilon = 0.1$ row is floor-dominated and uninformative despite its CI excluding zero; see discussion below.

ε	Race (White=0, Non-White=1)			Sex (Male=0, Female=1)		
	DP-Gap	δ_W/δ_{NW}	Δ_{PCER}	DP-Gap	δ_M/δ_F	Δ_{PCER}
no-DP	.058±.009	.0065/.0056	−13±180 [−62, +34]	.097±.012	.0064/.0054	+15±187 [−36, +66]
0.1	.033±.026	.0040/.0038	+12±220 [−49, +72]	.079±.035	.0039/.0034	+2±185 [−49, +53]
0.5	.038±.008	.0032/.0035	+14±210 [−43, +72]	.094±.010	.0035/.0030	+33±204 [−23, +90]
1.0	.039±.007	.0037/.0033	−14±218 [−71, +47]	.095±.008	.0035/.0034	+30±192 [−19, +83]
5.0	.041±.007	.0037/.0034	+3±222 [−56, +65]	.096±.008	.0036/.0030	+5±203 [−51, +61]
10.0	.042±.007	.0037/.0034	−7±217 [−64, +52]	.096±.008	.0037/.0031	+13±205 [−42, +70]

Table 2: ACS Income 2018 ($n = 50,000$, CA+TX+NY), mean $\pm$ std over 50 seeds. **Left half (Race)**: White=0, Non-White=1; positive prediction = income $> \$50K$ (beneficial). **Right half (Sex)**: Male=0, Female=1. $\delta_{g0/g1}$: per-group overfitting gap means ($g0/g1$); Δ_{PCER} : 95% bootstrap CI in brackets. All CIs cross zero; no robust PCER finding observed for ACS at any ε (sex direction is positive but not significant).

only marginally better than guessing, the well-documented weakness of shadow-model-free threshold attacks relative to calibrated attacks such as LiRA (Carlini et al. 2022). Despite this, δ_g correlates with $AUC_g - 0.5$ at $r = 0.828$ ($p < 10^{-18}$), consistent with the Yeom bound holding with room to spare (mean advantage $0.0013 \ll \text{mean } \delta_g \approx 0.005$). At the PCER-disparity level, Δ_{PCER}^δ and Δ_{PCER}^{MIA} agree in sign in 32/36 configurations (88.9%; 26/30 excluding $\varepsilon = 0.1$) and correlate at $r = 0.935$ ($p < 10^{-16}$). All four disagreements occur on ACS, precisely the combinations Section 5.3 already flags as noise-dominated: the direct-MIA comparison reproduces, rather than adds, instability.

The proxy is conservative in most configurations. Remark 2 bounds each group’s PCER value individually; it does not guarantee $|\Delta_{PCER}^\delta| \leq |\Delta_{PCER}^{MIA}|$ at the disparity level (Remark 2). We therefore test this pattern directly: it holds empirically in 69% of configurations (median ratio 1.25 where it holds), consistent with δ_g generally understating rather than overstating the disparity a direct attack would report, though not universally so. Together these results support using δ_g as the audit statistic: raw attack signal is too close to chance to serve as a stable denominator here, while the overfitting gap is stable, shadow-model-free, and tracks the direct attack’s PCER ranking in the large majority of informative settings.

6 Discussion

Why a proxy instead of a direct attack. Section 5.5 showed a real per-group threshold attack is only marginally

better than chance here (mean AUC advantage 0.0013), yet still agrees directionally with δ_g -based PCER in most settings, which is the practical case for the proxy: cheaper and less noisy than the attack it bounds. Practitioners with budget for calibrated attacks such as LiRA (Carlini et al. 2022) can substitute a tighter estimate directly into the denominator; Algorithm 1 is agnostic to which cost proxy is used.

The apparent fairness of high DP noise. Deploying $\varepsilon = 0.1$ because it minimizes DP-Gap can be misleading: the metric improves not because the system is more equitable, but because heavy noise compresses outcome rates toward uniformity, and PCER reduces to scaled DP-Gap there. The genuine question is what happens at moderate budgets where both outcomes and privacy risk remain distinguishable.

Privacy cost as a legitimate, but contestable, fairness input. The norm that involuntary privacy costs should be proportional to prediction benefits is grounded in compensatory fairness, not voluntary exchange, and is one motivated position rather than a conclusion forced by the data (Section 3.2 contrasts it with equality-of-treatment, distributive-justice, and Rawlsian alternatives). Under this norm, groups bearing greater involuntary risk who also receive fewer benefits are doubly disadvantaged, even when not disadvantaged by an equality-of-treatment standard. Neither NIST SP 800-226 (National Institute of Standards and Technology 2025) nor the EU AI Act’s Article 10 bias-monitoring requirements (European Parliament and Council of the European Union 2024) address this distributional dimension of privacy cost explicitly; PCER operationalizes a concrete audit

ε	Acc	DP-Gap	EO-Gap	δ_M	δ_F	Δ_{PCER}
no-DP	.960 $\pm$.001	.523 $\pm$.006	.070 $\pm$.012	.0292 $\pm$.0032	.0446 $\pm$.0018	+22 $\pm$ 3 [+21, +23]
0.1	.560 $\pm$.031	.079 $\pm$.042	.078 $\pm$.042	.0038 $\pm$.006	.0035 $\pm$.005	+65 $\pm$ 316 [-17, +152]
0.5	.912 $\pm$.005	.485 $\pm$.013	.092 $\pm$.023	.0036 $\pm$.004	.0054 $\pm$.004	+349 $\pm$ 347 [+255, +444]
1.0	.944 $\pm$.003	.514 $\pm$.008	.087 $\pm$.016	.0051 $\pm$.004	.0065 $\pm$.004	+236 $\pm$ 289 [+159, +317]
5.0	.961 $\pm$.002	.526 $\pm$.007	.078 $\pm$.014	.0053 $\pm$.003	.0075 $\pm$.003	+219 $\pm$ 243 [+155, +290]
10.0	.963 $\pm$.002	.526 $\pm$.006	.075 $\pm$.014	.0053 $\pm$.003	.0077 $\pm$.003	+209 $\pm$ 246 [+147, +279]

Table 3: Bias in Bios, gender attribute (Male=0, Female=1). Mean $\pm$ std over 50 independent runs (seeds 0–49). δ_g : per-group overfitting gap; Δ_{PCER} : PCER disparity (95% bootstrap CI in brackets). Positive prediction = professor ($y=1$, **beneficial**). DP-Gap exceeds 0.48 at all $\varepsilon \geq 0.5$; PCER is positive and CI excludes zero at no-DP and at $\varepsilon \geq 0.5$.

that both leave implicit, deployable with only train/test accuracy per group (Algorithm 1) and adding negligible overhead where it is informative ($\varepsilon \geq 1$, floor-robust at $\eta = 10^{-2}$), whichever normative standard a practitioner ultimately applies to the resulting measurement.

Relationship between DP-Gap, EO-Gap, and PCER.

DP-Gap and EO-Gap move together in most configurations here, while PCER diverges through the overfitting gap term. The Adult (race) ranking disagreement at the default floor ($\varepsilon \in \{5, 10\}$ best by DP-Gap/EO-Gap, no-DP best by PCER) illustrates this divergence, though it is a floor-clamping effect rather than a stable exposure signal: both groups’ overfitting gaps shrink by a similar proportion under DP-SGD (not more evenly *across* groups), which amplifies the existing positive-rate gap in the PCER denominator; raising the floor to $\eta = 10^{-2}$ reverses the ranking back into agreement with DP-Gap (see the extended version).

Structural impossibility under DP-SGD. DP-SGD provides no mechanism-level control over the between-group overfitting differential (Section 3.2). Consequently, whenever $\delta_0 \neq \delta_1$ (the typical case), Proposition 3 makes passing both a DP-Gap and a PCER audit structurally impossible: any system claiming both is implicitly asserting equal per-group overfitting.

Per-group overfitting as a proxy for unequal exposure. δ_g is an *upper bound* on $\hat{\mu}_g$, not a measurement of realized attack success, so $\delta_0 \neq \delta_1$ formally shows only that our worst-case bound differs across groups. We nonetheless interpret persistent $\delta_0 \neq \delta_1$ as suggestive evidence of unequal effective privacy protection under the shared nominal ε : overfitting is the mechanism through which DP-SGD’s noise budget fails to protect a group uniformly, a reading the direct-attack validation above supports without formally establishing.

Domain-sensitive benefit interpretation. In harmful-prediction domains such as COMPAS, PCER’s ratio structure is unchanged, but it now captures a *harm-to-cost* ratio rather than a benefit-to-cost one, so the normative interpretation of Δ_{PCER} ’s sign is domain-specific.

Measurement limitations and extensions. Using R_g as a benefit proxy conflates base-rate differences with model-induced disparities (see Limitations). Δ_{PCER} ’s absolute magnitude depends on both the positive-rate gap and the overfitting level and is not comparable across datasets; practitioners

should focus on *sign* and *floor-robustness* rather than magnitude in isolation. A natural dimensionless normalization would divide Δ_{PCER} by the mean PCER, which we leave to future calibration work. PCER as defined also treats the protected attribute as binary; extending to multi-class attributes (age brackets, multi-category race) is straightforward at the auditor level via a worst-case dispersion measure such as $\max_{g, g'} |\text{PCER}_g - \text{PCER}_{g'}|$, requiring no change to the trainer’s released statistics.

Although motivated by DP-SGD, PCER depends only on δ_g and R_g , both computable for any training regime, so Algorithm 1 applies equally to federated learning, PATE (Papernot et al. 2017), or unprotected ERM. Practitioners outside the DP-SGD setting can therefore apply PCER as a general post-hoc equity audit without modification.

Affected communities. The communities represented in these benchmarks (see the Ethics Statement) bear the highest stakes of the questions we raise. Definitions of “benefit” and “proportional compensation” should ultimately be co-developed with affected communities in specific deployment contexts; the PCER metric is a diagnostic tool, not a substitute for that engagement.

7 Conclusion

We have proposed the Privacy-Cost Equity Ratio (PCER), a group fairness criterion operationalizing harm-proportionality using Adams’ equity ratio structure: the ratio of prediction benefit to involuntary privacy exposure (proxied by δ_g), grounded in the Yeom et al. bound ($\text{Adv}_{\text{MIA}}(g) \leq \delta_g$) and requiring only train/test accuracy per group. Across Adult Income, COMPAS, ACS Income 2018, and Bias in Bios (six benchmark-attribute combinations) under DP-SGD at moderate budgets ($\varepsilon \geq 1$), PCER diverges from demographic parity gap in ways whose *sign* is robust to floor-constant choice, capturing on COMPAS a persistent lower benefit-to-cost ratio for African-Americans that DP-Gap misses entirely. A direct per-group threshold membership-inference attack, run on every trained model as a validation check (Section 5.5), agrees in sign with the overfitting-gap proxy in 32/36 configurations and correlates with it at $r = 0.935$, with all four disagreements confined to ACS, which the sensitivity analysis already flags as noise-dominated. Algorithm 1 packages the procedure as a deployable post-hoc audit requiring no auxiliary infrastructure.

A critical caveat: PCER degenerates under strong DP ($\epsilon = 0.1$), where overfitting gaps collapse toward zero and Δ_{PCER} reduces to a scaled DP-Gap with no additional information. The apparent directional instability on COMPAS at $\epsilon = 0.1$ is exactly this floor artifact. Practitioners should apply PCER only at DP budgets where group overfitting is reliably above the floor, verifying with the floor-sensitivity table in the extended version.

7.1 Limitations

Six constraints bound the conclusions above.

1. **Floor-Dominated Regime.** PCER is uninformative when both groups' overfitting gaps are near zero ($\epsilon = 0.1$ in these experiments). Results are reported at $\eta = 10^{-3}$ throughout (full sensitivity analysis in the extended version), and any finding at strong DP budgets requires floor-sensitivity verification before interpretation. We recommend treating only findings that are stable across at least two decades of η as reliable, since unstable findings are likely floor artifacts rather than genuine privacy-cost inequity.
2. **Statistical Power.** Fifty seeds provide reasonable precision for estimating mean and standard deviation of PCER, including on COMPAS ($n = 6,172$) where within-group sample sizes are small. Pearson correlations between PCER and DP-Gap are computed from five ϵ settings and should be treated as descriptive summaries rather than inferential tests; we report bootstrap confidence intervals (5,000 resamples, percentile method) throughout. Larger seed counts would tighten these intervals further, particularly at the floor-dominated $\epsilon = 0.1$ setting where per-seed variance is highest, but the qualitative floor-versus-robust distinction we draw does not depend on this precision improving, since it rests on the sign and order of magnitude of Δ_{PCER} rather than on tight interval widths.
3. **Overfitting Gap as Privacy Proxy.** The Yeom bound $\text{Adv}_{\text{MIA}}(g) \leq \delta_g$ is an upper bound; the true MIA advantage may be substantially lower than δ_g when the model generalizes well. Using δ_g as the denominator of PCER therefore *overestimates* each group's privacy cost and understates its individual PCER value (Remark 2); this per-group conservativeness does not guarantee the reported *disparity* Δ_{PCER} is itself conservative, so its magnitude should be read alongside the direct-attack validation (Section 5.5), where it holds in 69% of configurations rather than universally. For settings with non-trivial memorisation, LiRA (Carlini et al. 2022) or similar attacks could provide tighter per-group advantage estimates.
4. **Positive Rate as Benefit Proxy.** R_g reflects base-rate differences as well as model-induced disparities. Studies with access to qualification or eligibility covariates could construct a more principled benefit measure, potentially separating group-level differences in benefit from pre-existing distributional differences in the data.
5. **Dataset and Architecture Scope.** Our benchmarks use a simple MLP on tabular inputs and TF-IDF+SVD NLP features. The Bias in Bios results show that PCER findings replicate across modalities when features are fixed offline,

but they may not generalize to end-to-end fine-tuned language models where DP-SGD interacts with large-scale parameter memorisation. Extending PCER evaluation to neural language models under DP fine-tuning is an important open question.

6. **Fixed Hyperparameters Across Budgets.** All DP-SGD runs share a single learning rate, clipping norm, batch size, and epoch count (Section 4) rather than a per- ϵ tuned configuration; noise, learning rate, and clipping norm interact non-trivially in DP-SGD, so a jointly tuned sweep could shift the per-group overfitting gaps δ_g reports and, with them, the specific PCER estimates above. We expect the qualitative floor-sensitivity pattern to persist regardless, since it follows from Proposition 2's degeneracy argument rather than from any particular hyperparameter choice, but the exact budgets at which PCER estimates stabilize could shift under a per-budget tuned sweep. This is standard practice in DP-SGD auditing work, which typically fixes a single training recipe across the privacy-budget sweep to isolate the effect of ϵ itself, but it means the absolute PCER values reported here should be read as characterizing this particular training recipe rather than DP-SGD in general. Auditors adopting PCER for a different architecture or optimizer configuration should treat the floor constant η and the specific budgets at which findings stabilize as recipe-dependent quantities to be re-derived, not as universal constants transferable from this study.

Future work. Three extensions are immediate. First, individual-level PCER using per-sample overfitting or MIA vulnerability scores would recover Adams' original individual-to-referent comparison; the natural instrument here is a per-example loss trajectory across training checkpoints, since per-sample overfitting need not track the group mean uniformly, and outlier-driven exposure could be masked by our current group-level averaging. Second, a mechanism-design extension: training DP models to jointly minimize PCER disparity and DP-Gap is a natural objective for constrained DP learning, for instance via a Lagrangian penalty on Δ_{PCER} analogous to existing fair-DP-SGD approaches (Tran, Fioretto, and Van Hentenryck 2021) but targeting the benefit-to-cost ratio rather than an outcome-rate gap alone; whether such a penalty can be enforced without destabilizing the privacy accountant is an open question. Third, replacing the raw positive rate R_g with a calibration-adjusted benefit measure would remove the base-rate confound and make PCER more directly diagnostic of model-induced inequity, for example by conditioning on predicted-probability calibration within each group before computing R_g , so that residual disparities reflect genuine model behavior rather than pre-existing label-rate differences across groups. We view these three directions as complementary rather than competing: the first refines the unit of analysis, the second turns the audit into a training-time constraint, and the third sharpens the benefit side of the ratio, so a deployment-ready successor to PCER could plausibly combine all three, and each extends Algorithm 1 without changing its two-part trainer/auditor structure, so none requires re-architecting the audit pipeline itself.

Ethics Statement

This paper analyzes publicly available benchmark datasets (COMPAS, Adult Income, ACS Income, and Bias in Bios) that encode sensitive demographic attributes of individuals who did not consent to be subjects of algorithmic fairness research. COMPAS in particular records recidivism risk scores that influenced real pre-trial detention and sentencing decisions, over populations disproportionately Black and with limited capacity to contest algorithmic assessments of their futures; reusing these datasets, even for audit purposes, reproduces and circulates records of consequential harm.

No new data collection or human-subjects research was conducted and IRB review was not required. Beyond the formal requirements, we note that framing privacy exposure as a cost implicitly treats membership inference risk as a background condition of currently deployed systems, rather than a harm that could in principle be eliminated. This is a normative choice reflecting the constraints of existing deployed systems, rather than an endorsement of them.

PCER is a diagnostic tool, and we have tried to be precise about what it does and does not measure, but it could be misused once it leaves our hands. A deployer could cite a favorable PCER reading at one floor constant while omitting the floor-sensitivity analysis revealing the finding as unstable; our robustness check is only advisory, and nothing prevents selective reporting.

There is also a subtler risk that privacy exposure can be framed as a cost that partially offsets outcome disparity. This could be appropriated to argue a group receiving fewer positive predictions is compensated by lower membership inference risk; in the paper, we reject that interpretation but cannot prevent it. Publishing a quantitative audit methodology also risks compliance theater, where a single score substitutes for structural accountability, as aggregate metrics can deflect substantive demands. PCER should always be reported alongside the floor-sensitivity table, bootstrap CIs, and a plain-language account for affected communities; qualitative and regulatory review remain necessary.

The authors come to this work from a computer science and machine learning background, with training in formal fairness metrics, differential privacy theory, and empirical evaluation, rather than firsthand experience of the harms encoded in these datasets or the communities most affected by criminal-justice risk scoring or income prediction. Grounding our criterion in Adams' equity theory reflects a tradition within Western organizational psychology that may not resonate with community-centered or reparative conceptions of justice; likewise, operationalizing privacy cost via the overfitting gap and binary demographic groups reflects pragmatic constraints on scalable auditing, rather than claims about the basis of demographic categories or the sufficiency of group-level analysis. Researchers with different backgrounds or lived experience might have asked different questions, or contested this framing before building on it.

Acknowledgements

An AI coding assistant helped implement experiment code and L^AT_EX; all research design, analysis, interpretation, and

writing reflect the authors' independent judgment.

References

- Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 308–318. ACM.
- Adams, J. S. 1963. Towards an Understanding of Inequity. *Journal of Abnormal and Social Psychology*, 67(5): 422–436.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. [Accessed 22-05-2026].
- Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 15479–15488. Curran Associates.
- Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, 149–159. PMLR.
- Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramèr, F. 2022. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (S&P)*, 1897–1914. IEEE.
- Chang, H.; and Shokri, R. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 292–303. IEEE.
- Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 309–315. ACM.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*, 120–128. ACM.
- Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 6478–6490. Curran Associates.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science (ITCS)*, 214–226. ACM.
- Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)*, 265–284. Springer.
- European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union. Article 10.

- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 3315–3323. Curran Associates.
- Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; and Ullman, J. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3000–3008. PMLR.
- Kulynych, B.; Yaghini, M.; Cherubin, G.; Veale, M.; and Troncoso, C. 2022. Disparate Vulnerability to Membership Inference Attacks. *Proceedings on Privacy Enhancing Technologies*, 2022(1): 460–480.
- National Institute of Standards and Technology. 2025. Guidelines for Evaluating Differential Privacy Guarantees. Technical Report SP 800-226, NIST.
- Nussbaum, M. C. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA and London, England: Harvard University Press.
- Papernot, N.; Abadi, M.; Erlingsson, Ú.; Goodfellow, I. J.; and Talwar, K. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *International Conference on Learning Representations (ICLR)*.
- Romanov, A.; De-Arteaga, M.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; Rumshisky, A.; and Kalai, A. T. 2019. What’s in a Name? Reducing Bias in Bios Without Access to Protected Attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4187–4195. ACL.
- Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (S&P)*, 3–18. IEEE.
- Speicher, T.; Heidari, H.; Grgic-Hlaca, N.; Gummadi, K. P.; Singla, A.; Weller, A.; and Zafar, M. B. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2239–2248. ACM.
- Tran, C.; Fioretto, F.; and Van Hentenryck, P. 2021. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 9932–9939. AAAI Press.
- Uniyal, A.; Naidu, R.; Kotti, S.; Singh, S.; Kenfack, P. J.; Mireshghallah, F.; and Trask, A. 2021. DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy? arXiv:2106.12576.
- Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. IEEE.
- Yousefpour, A.; Shilov, I.; Sablayrolles, A.; Testuggine, D.; Prasad, K.; Malek, M.; Nguyen, J.; Ghosh, S.; Bharadwaj, A.; Zhao, J.; Cormode, G.; and Mironov, I. 2021. Opacus: User-Friendly Differential Privacy Library in PyTorch. arXiv:2109.12298.