A CLOSER LOOK AT THE CALIBRATION OF DIFFERENTIALLY PRIVATE LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We systematically study the calibration of classifiers trained with differentially private stochastic gradient descent (DP-SGD) and observe miscalibration across a wide range of vision and language tasks. Our analysis identifies per-example gradient clipping in DP-SGD as a major cause of miscalibration, and we show that existing baselines for improving private calibration only provide small improvements in calibration error while occasionally causing large degradation in accuracy. As a solution, we show that differentially private variants of post-processing calibration methods such as temperature calibration and Platt scaling are surprisingly effective and have negligible utility cost to the overall model. Across 7 tasks, temperature calibration and Platt scaling with DP-SGD result in an average 58-fold reduction in the expected calibration error and only incurs an up to 1.59 percent drop in accuracy.

1 INTRODUCTION

Modern deep learning models tend to memorize their training data in order to generalize better (Zhang et al., 2021; Feldman, 2020), posing great privacy challenges including training data leakage and membership inference attacks (Shokri et al., 2017; Hayes et al., 2017; Carlini et al., 2021). To address these concerns, differential privacy (DP) has become a popular paradigm for providing rigorous privacy guarantees when performing data analysis and statistical modeling based on private data. In practice, DP-SGD algorithm (Abadi et al., 2016) is proposed to train neural networks with provably limited privacy leakage, which involves the combination of per-example gradient clipping and gradient noise injection during the optimization process.

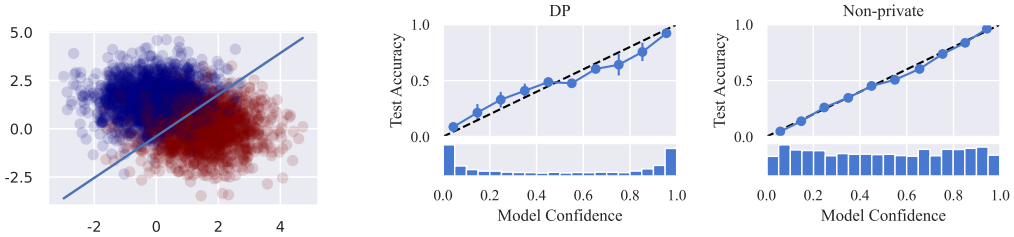
Despite the usefulness of DP-SGD in ensuring user privacy, prior works have identified that this privacy comes at the cost of other aspects of trustworthy machine learning, such as lower accuracy and fairness (Bagdasaryan et al., 2019; Feldman, 2020; Sanyal et al., 2022). These tradeoffs pose a challenge for privacy-preserving machine learning, as it forces practitioners to make difficult decisions on how to weigh privacy against other key aspects of trustworthiness. In this work, we expand the study of privacy-related tradeoffs by characterizing and proposing mitigations for the *privacy-calibration* tradeoff.

The existence of such a tradeoff may be surprising, as we might expect differentially private training to *improve* calibration by preventing models from memorizing training examples and ensuring test-set generalization (Bassily et al., 2016; Kulynych et al., 2022). However, we find that DP training has the surprising effect of consistently producing over-confident prediction scores (Bu et al., 2021) in practice. We show an example of this phenomenon in a simple 2D logistic regression problem over Gaussian data (Fig. 1). We find a polarization phenomenon, where the DP-trained model achieves similar accuracy to its non-private counterpart, but its confidences are clustered around either 0 or 1. As we will see later, the polarization insight conveyed by this motivating example transfer to more realistic settings.

Our first contribution quantifies existing privacy-calibration tradeoffs on state-of-the-art models that leverage differentially private training and pre-trained backbones such as RoBERTa (Liu et al., 2019)

and vision transformers (ViT) (Dosovitskiy et al., 2020). While there have been some studies of miscalibration for differentially private learning (Bu et al., 2021; Knolle et al., 2021), they focus on simple tasks (MNIST, SNLI) with small networks trained from scratch. Our work shows the miscalibration problems persist even with state-of-the-art private models that nearly match non-private accuracies. Through controlled experiments and ablations, we show that these calibration errors are unlikely to be due to the accuracy or regularization effects of DP-SGD, and are likely caused by the per-example clipping operation in DP-SGD.

Our second contribution shows that this tradeoff can be easily addressed through differentially private variants of temperature scaling (DP-TS) and Platt scaling (DP-PS). To enable these modifications, we provide a simple privacy accounting analysis, proving that DP-SGD based recalibration on a held-out split does not incur additional privacy costs. Through extensive experiments, DP-TS and DP-PS effectively prevent overconfidence of DP-trained models and reduces calibration error by up to $7.5\times$, substantially outperforming more complex interventions that have been claimed to improve calibration (Bu et al., 2021; Knolle et al., 2021).



(a) Non-separable Gaussian Data

(b) Calibration comparison of logistic regression w and w/o DP training

Figure 1: (a) Logistic Regression model (blue line) with $\epsilon = 8$ on Gaussian data $\{(x_i, y_i)\}_{i=1}^n$ where $(x, y) \in \mathbb{R}^p \times \{1, -1\}, (x - b)|y \sim \mathcal{N}(0, I_{2\times 2}), b = [1.5, 0]$ if $y = 1$ else $b = [0, 1.5]$. (b) DP-SGD trained classifier, which shows poor calibration with a large concentration of extreme confidence values (**Left**); the baseline is a standard, non-private logistic regression model trained by SGD, which is much better calibrated (**Right**).

2 Related Work

Differentially Private Deep Learning. DP-SGD (Song et al., 2013; Abadi et al., 2016) is a popular way to guarantee privacy for deep learning (DL) models and involves clipping per-example gradients and adding gradient noise. Recent works have shown that fine-tuning high-quality pre-trained models with DP-SGD produces high downstream performance (Tramer & Boneh, 2021; Li et al., 2022; De et al., 2022). Existing works have studied how ensuring differential privacy through mechanisms such as DP-SGD leads to tradeoffs with other properties, such as accuracy (Feldman, 2020) and fairness (Bagdasaryan et al., 2019; Sanyal et al., 2022) (as measured by accuracy on minority groups). Our work expands our understanding of tradeoffs between privacy and other aspects of trustworthiness by characterizing privacy-calibration tradeoffs.

Calibration. Calibrated probability estimates match the true empirical frequencies of an outcome, and calibration is often used to evaluate the quality of uncertainty estimates provided by neural networks. Recent works have observed that highly-accurate models that leverage pre-training are often well-calibrated (Desai & Durrett, 2020; Minderer et al., 2021; Kadavath et al., 2022). However, we find that even pre-trained models are poorly calibrated when they are fine-tuned using DP-SGD.

Our work is not the first to study calibration under learning with differential privacy, but we provide a more comprehensive characterization of privacy-calibration tradeoffs and solutions that improve this tradeoff which are both simpler and more effective. Luo et al. (2020) study private calibration for out-of-domain situations, but do not study whether DP-SGD causes miscalibration in-domain. Angelopoulos et al. (2021) modifies split conformal prediction to be privacy-preserving, but they only study vision models and their private models have substantial performance decrease compared

to non-private ones. They also do not study the causes of the *privacy-calibration* tradeoff. Knolle et al. (2021) study miscalibration, but only on MNIST and a small pneumonia dataset. Our work provides a more comprehensive characterization across more realistic datasets, and our comparisons show that our proposed recalibration approach is consistently more effective. Bu et al. (2021) is the closest work to ours. They identify that DP-SGD is miscalibrated on CIFAR-10, SNLI, and MNIST. As a solution, they suggest an alternative clipping scheme that reduces expected calibration error (ECE). Our work differs in three ways: our experimental results cover harder tasks and control for confounders such as model accuracy and regularization; we study a pre-training transfer setting that is closer to the state-of-the-art in differentially private learning and find substantially worse ECE gaps (e.g. they identify a 43% relative increase in ECE on CIFAR-10, while we find nearly 400% on Food101); and finally, we compare our simple recalibration procedure to their method and find that private temperature scaling is substantially more effective at reducing ECE than the proposed global clipping technique.

3 Problem Statement and Methods

One of our goals in this work is to build classifiers that are both accurate and calibrated under differential privacy. We begin by defining core concepts.

3.1 Differential Privacy

Differential privacy is a formal privacy guarantee for a randomized estimator that ensures that no adversary has a high probability of identifying whether a record was included in a dataset X.

Definition 3.1. (Approximate-DP (Dwork et al., 2006)). The randomized algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-DP if for all neighboring datasets $X, X' \in \mathcal{X}$ that differ on a single element and all measurable $Y \subset \mathcal{Y}, \mathbb{P}(\mathcal{M}(X) \in Y) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{M}\left(X'\right) \in Y\right) + \delta$.

The privacy budget ϵ bounds the log-likelihood ratio between the output distribution of a randomized algorithm $\mathcal{M}$ run on a dataset X and the output when it is run on X'. The failure probability δ allows the guarantee to hold approximately. Throughout our work, we will study models trained with (ϵ, δ)-DP algorithms.

3.2 Differentially Private Stochastic Gradient Descent

The standard approach to train neural networks with differential privacy is through the use of the differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016) algorithm. The algorithm operates by privatizing each gradient update via combining per-example gradient clipping and Gaussian noise injection.

Formally, one step of DP-SGD to update θ with a batch of samples $\mathcal{B}_t$ involves the following process

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \left\{ \frac{1}{B} \sum_{i \in \mathcal{B}_t} \text{clip}_C \left(\nabla \mathcal{L}_i \left(\theta^{(t)} \right) \right) + \xi \right\}, \tag{1}$$

where η_t is the learning rate at step t, $\text{clip}_C \left(\nabla l_i \left(w^{(t)} \right) \right)$ clips the gradient by performing $\text{clip}_C \left(\nabla l_i \left(w^{(t)} \right) \right) = \nabla \mathcal{L}_i \left(w^{(t)} \right) \cdot \min \left(1, C / \| \nabla \mathcal{L}_i \left(w^{(t)} \right) \|_2 \right)$ and ξ is the Gaussian noise defined as $\xi \sim \mathcal{N} \left(0, C^2 \frac{\sigma^2}{B^2} I_p \right)$ with the standard deviation σ and expected batch size B. Each step of DP-SGD fulfills an (ϵ, δ)-DP guarantee, and the final model satisfies DP with privacy leakage parameters that can be computed with composition theorems Mironov (2017); Dong et al. (2019); Gopi et al. (2021).

3.3 Calibration

A probabilistic forecaster is said to be *calibrated* if the forecaster has accuracy p on the set of all examples with confidence p. Specifically, we say that a probabilistic classifier h over C classes satisfies *canonical calibration* if for all p in the simplex Δ^{C-1} and label y, $P(Y = y \mid h_\theta(X) = p) = p_y$ holds.[1]

[1]We slightly abuse the notation of X and Y.

The canonical calibration property can be difficult to verify in practice when the number of classes is large. Because of this, we will consider a simpler top-label calibration criterion in this work. In this relaxation, we consider calibration over only the highest probability class. More formally, we say that a classifier h is calibrated if

$$\forall p^* \in [0,1], P\left(Y \in \arg\max p \mid \max h_\theta(X) = p^*\right) = p^*. \tag{2}$$

We will quantify the degree to which a classifier is calibrated through the expected calibration error (ECE), defined by

$$\mathbb{E}\left[\left|\, p^* - E\left[Y \in \arg\max h_\theta(X) \left|\max h_\theta(X) = p^*\right.\right]\right|\right]. \tag{3}$$

In practice, we estimate ECE by first partitioning the confidence scores into B_m bins and then calculating the empirical estimate of ECE as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}\left(B_m\right) - \text{conf}\left(B_m\right)\right|, \tag{4}$$

where $\text{acc}\left(B_m\right) = \frac{1}{|B_m|}\sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i = y_i\right)$ and $\text{conf}\left(B_m\right) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i$. When appropriate, we will also study fine-grained miscalibration errors through the histogram of $\text{conf}(B_m)$ (confidence histogram) and plot $\text{acc}(B_m)$ against $\text{conf}(B_m)$ (the reliability diagram).

3.4 Recalibration

When models are miscalibrated, *post-hoc recalibration* methods are often used to reduce calibration errors. These methods reduce miscalibration by adjusting the log-probability scores generated by the probabilistic prediction model. More formally, consider a score-based classifier that produces probabilistic forecasts via $\text{softmax}(h_\theta(\mathbf{x}))$. We can adjust the calibration of this classifier by learning a g_ϕ that adjusts the log probabilities and produces a better calibrated forecast $\text{softmax}(g_\phi(h_\theta(\mathbf{x})))$.

Typically, the re-calibration function g is learned by minimizing a proper scoring rule on a separate validation set with the following optimization problem

$$\min_\phi \mathbb{E}[\ell(\text{softmax}(g_\phi \circ h_\theta(\mathbf{x})), y)]. \tag{5}$$

Specific examples of this type of post-hoc recalibration technique includes temperature scaling (Guo et al., 2017), Platt scaling (Platt et al., 1999), and isotonic regression (Zadrozny & Elkan, 2002). These methods differ in their choice of g_ϕ, and in this work we will consider temperature scaling ($g_\phi(\mathbf{x}) = \mathbf{x}/T$) and Platt scaling ($g_\phi(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}$) with the choice of log-loss for ℓ.

3.5 Differentially private recalibration

As observed in our motivating example (Fig. 1) and later experimental results, DP training (with DP-SGD and DP-Adam) tends to produce miscalibrated models. Motivated by the success of recalibration methods such as Platt scaling and temperature scaling in the non-private setting (Guo et al., 2017), we study how these methods can be adapted to build well-calibrated classifiers with a differential privacy guarantee.

Our proposed approach is extremely simple and consists of three steps, shown in Algorithm 1. We first split the training set into a model training part (X_{train}) and recalibration part (X_{recal}). We then train the model using DP-SGD, followed by a recalibration step using DP-SGD. Depending on the choice of g_ϕ, we will refer to this algorithm as either DP temperature scaling (DP-TS) or DP Platt scaling (DP-PS).

The use of sample splitting for recalibration makes privacy accounting extremely simple. To achieve a target (ϵ, δ)-DP guarantee after recalibration, we can simply run both stages with DP-SGD parameters that achieve (ϵ, δ)-DP (Prop. A.1). While sample splitting does reduce the number of samples available for model training, using 90% of the dataset for X_{train} results in minimal utility cost for the model training step in practice.

Algorithm 1: Differentially private recalibration

Input: $X = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)\}$, validation ratio α.
Initial: Parameters of models h_θ, recalibrator g_ϕ.

1. $X_{\text{train}}, X_{\text{recal}} = \text{RandomSplit}(X, \alpha)$
2. Train $h_\theta(\mathbf{x})$ using DP-SGD to optimize $\min_\theta \mathbb{E}\left[\ell\left(\text{softmax}\left(h_\theta(\mathbf{x})\right), y\right)\right]$ with X_{train}
3. Train g_ϕ using DP-SGD to optimize $\min_\phi \mathbb{E}\left[\ell\left(\text{softmax}\left(g_\phi \circ h_\theta(\mathbf{x})\right), y\right)\right]$ with X_{recal}

Output: $g_\phi \circ h_\theta(\cdot)$

4 Experimental Results

We study three different experimental settings. In each case, we will show that DP-SGD leads to high miscalibration, and DP recalibration substantially reduces calibration error.

We first consider **in-domain** evaluations, where we evaluate calibration errors on the same domain that they are trained on. Results show that pre-trained models do not address miscalibration issues in-domain. We then evaluate the same models above in **out-of-domain** settings, showing that both miscalibration and effectiveness of our re-calibration methods carry over to the out-of-domain setting. Finally, we perform careful **ablations** to isolate and understand the causes of in-domain miscalibration.

Models Our goal is to evaluate calibration errors for state-of-the-art private models. Because of this, our models are based on transfer learning from a pre-trained model. For the text datasets, we perform fine-tuning for RoBERTa-base using the procedure of Li et al. (2022), and for vision datasets, we perform linear probing of ViT and ResNet-50 features, analogous to Tramer & Boneh (2021).

Datasets Following prior work (Li et al., 2022), we train on MNLI, QNLI, QQP, SST-2 (Wang et al., 2019) for the text tasks, and perform OOD evaluations on common transfer targets such as Scitail (Khot et al., 2018), HANS (McCoy et al., 2019), RTE, WNLI, and MRPC (Wang et al., 2019).[2] For the vision tasks, we focus on the in-domain setting and evaluate on a subset of the transfer tasks in Kornblith et al. (2019) with at least 50k examples.

Methods As baselines, we trained the above models using non-private SGD (Non-private), standard DP-SGD (DP), global clipping (Bu et al., 2021) (Global Clipping), and differentially private stochastic gradient Langevin dynamics (Knolle et al., 2021) (DP-SGLD). These last two methods were included to evaluate our simple recalibration approach against existing methods which were reported to improve calibration.

For our recalibration methods, we run the private recalibration method over the in-domain validation set X_{recal} in Sec. 3.5 using private temperature scaling (DP-TS) (Guo et al., 2017) and Platt scaling (DP-PS) (Platt et al., 1999; Guo et al., 2017). We also include a non-private baseline that combines differentially private model training with non-private temperature scaling (DP+Non-private-TS) as a way to quantify the privacy costs in the post-hoc recalibration step. Further implementation details and default hyper-parameters for DP training (Tab. 6) are in Appendix B.

4.1 In-domain Calibration

We now conduct in-depth experiments across multiple datasets and domains to study miscalibration (Tab. 1, 2). We train differentially private models using pre-trained backbones, and find that their accuracies match previously reported high performance (Tramer & Boneh, 2021; Li et al., 2022; De et al., 2022).

However, we find that these same models have substantially higher calibration errors. For example, linear probing for Food101 in Fig. 2 has private accuracy within 7 percent of the non-private

[2] To match the label space between MNLI and the OOD tasks, we merge "contradiction" and "neutral" labels into a single "not-contradiction" label.

Category	Model	CIFAR-10		SUN397		Food101	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.7951	0.0903	0.6844	0.1302	0.7582	0.154
	DP-SGLD	0.7122	0.1331	0.6062	0.1952	0.6476	0.2416
	Global Clipping	0.7712	0.0804	0.6215	0.1125	0.7451	0.1017
Recalibration	DP-PS	0.789	**0.012**	0.674	0.104	0.7543	0.0554
	DP-TS	0.789	0.0221	0.674	**0.0763**	0.7543	**0.0540**
Non-private	DP+Non-private-TS	0.789	0.0222	0.674	0.0764	0.7543	0.0539
	Non-private	0.83	0.0794	0.7044	0.1062	0.8245	0.0349

Table 1: The image classification performance ($\epsilon = 8$) of different models before and after recalibration. Results for $\epsilon = 3$ are in Appendix B.3

Category	Model	MNLI		QNLI		QQP		SST-2	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.826	0.1663	0.8464	0.1522	0.8659	0.1323	0.8842	0.114
	DP-SGLD	0.7188	0.2625	0.7787	0.2138	0.7917	0.2009	0.82	0.1742
	Global Clipping	0.8236	0.1667	0.8502	0.1491	0.8685	0.1296	0.8922	0.1047
Recalibration	DP-PS	0.826	**0.0487**	0.8464	**0.0305**	0.8659	0.0672	0.8842	**0.0201**
	DP-TS	0.826	0.0849	0.8464	0.0915	0.8659	**0.0635**	0.8842	0.0665
Non-private	DP+Non-private-TS	0.826	0.0849	0.8464	0.0915	0.8659	0.0635	0.8842	0.0665
	Non-private	0.8642	0.0699	0.914	0.028	0.9042	0.0891	0.9323	0.0425

Table 2: The text classification performance ($\epsilon = 8$) before and after recalibration.

counterpart, but the ECE is more than $4\times$ that of the non-private counterpart. In the language case, we see similar results on QNLI with a 6% drop in accuracy but $5\times$ increase in ECE. The overall trend of miscalibration is clear across datasets and modalities (Fig. 2).

DP recalibration. We now turn our attention to recalibration algorithms and see whether DP-TS and DP-PS can address in-domain miscalibration. We find that DP-TS and DP-PS perform consistently well over all datasets and on both modalities (Tab.1 and Tab.2). In many cases, the differentially private variants of recalibration work nearly as well as their *non-private* counterparts. The ECE values for the private DP-TS and non-private baseline of DP+Non-private-TS are generally close across all the datasets.

We note that both DP-TS and DP-PS perform consistently well, with an average relative (in-domain) ECE reduction of 0.58. Despite being simple, the two methods never underperform Global Clipping and DP-SGLD in terms of ECE, and generally have higher accuracies despite the added cost of sample splitting.

Qualitative analysis. Examining the reliability diagram before and after DP-TS, we see two clear phenomena. First, the model confidence distribution under DP-SGD is highly polarized (Fig. 3, first two panels) with nearly all examples receiving confidence of 1.0. Next, we see that after DP-TS, this confidence distribution is adjusted to cover a much broader range of confidence values. In the case of SUN397, we see almost perfect agreement between the model confidence and actual accuracies.

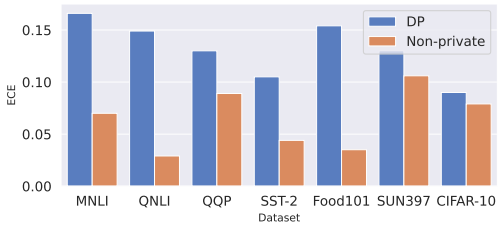


Figure 2: DP trained models display consistently higher ECE than their non-private counterparts.

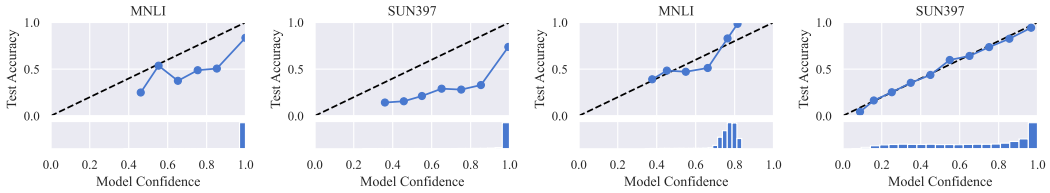


Figure 3: Reliability diagram before (**Left**) and after (**Right**) Recalibration using DP-TS. Recalibration parameters are learned on the validation set of MNLI and SUN397.

4.2 Out-of-domain Calibration

We complement our in-domain experiments with out-of-domain evaluations. To do this, we evaluate the zero-shot transfer performance of models trained over MNLI, QNLI (Tab. 3) and QQP (Tab. 4).

Our findings are consistent with the in-domain evaluations. Differentially private training generally increases ECE, and differentially private temperature scaling generally improves calibration. The gaps out-of-domain are substantially smaller than the in-domain case, as all methods are of low accuracy and miscalibrated out of domain. However, the general ranking of miscalibration methods, and the observation that DP-TS leads to private models with calibration errors on-par to non-private models is unchanged.

Dataset	Category	Model	Hans		Scitail		RTE		WNLI	
			Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
MNLI	Baseline	DP	0.5237	0.4752	0.7707	0.222	0.7220	0.2649	0.4366	0.5624
		DP-SGLD	0.4996	0.4995	0.7515	0.233	0.6498	0.3169	0.4507	0.5491
		Global Clipping	0.5	0.4747	0.4962	0.2051	0.4729	0.2737	0.5634	0.5632
	Recalibration	DP-PS	0.5237	**0.348**	0.7707	**0.1089**	0.7220	**0.1516**	0.4366	**0.4416**
		DP-TS	0.5237	0.3544	0.7707	0.1168	0.7220	0.1593	0.4366	0.4495
	Non-private	DP+Non-private-TS	0.5237	0.3544	0.7707	0.1168	0.7220	0.1593	0.4366	0.4495
		Non-private	0.668	0.2687	0.7853	0.1348	0.7906	0.1518	0.507	0.4677
QNLI	Baseline	DP	0.5002	0.4997	0.7377	0.26	0.5632	0.4439	0.4648	0.5332
		DP-SGLD	0.5	0.4986	0.7209	0.2723	0.5668	0.4266	0.4225	0.5738
		Global Clipping	0.5025	0.4971	0.7293	0.2684	0.5199	0.4761	0.4789	0.52
	Recalibration	DP-PS	0.5002	**0.3244**	0.7377	**0.0832**	0.5632	**0.2578**	0.4648	**0.3464**
		DP-TS	0.5002	0.385	0.7377	0.1353	0.5632	0.3121	0.4648	0.404
	Non-private	DP+Non-private-TS	0.5002	0.385	0.7377	0.1353	0.5632	0.3121	0.4648	0.404
		Non-private	0.538	0.1969	0.7454	0.0690	0.5199	0.3036	0.5493	0.2438

Table 3: The **zero-shot transfer** NLI performance ($\epsilon = 8$) across multiple OOD test datasets.

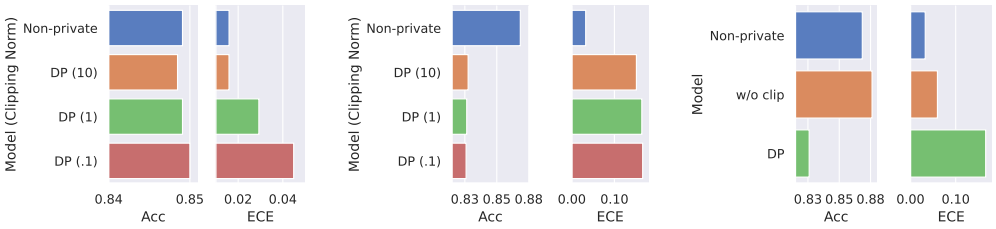
Dataset	Category	Model	MRPC	
			Accuracy	**ECE**
QQP	Baseline	DP	0.7426	0.2564
		DP-SGLD	0.6936	0.2979
		Global Clipping	0.7475	0.252
	Recalibration	DP-PS	0.7426	**0.1252**
		DP-TS	0.7426	0.1796
	Non-private	DP+Non-private-TS	0.7426	0.1796
		Non-private	0.7255	0.2635

Table 4: The **zero-shot transfer** paraphrase performance ($\epsilon = 8$) from QQP to MRPC.

4.3 Analysis and Ablation Studies

Finally, we carefully study two questions to better understand miscalibration of private learners: What component of DP-SGD leads to miscalibration? What are other confounders such as accuracy or regularization effects that lead to miscalibration?

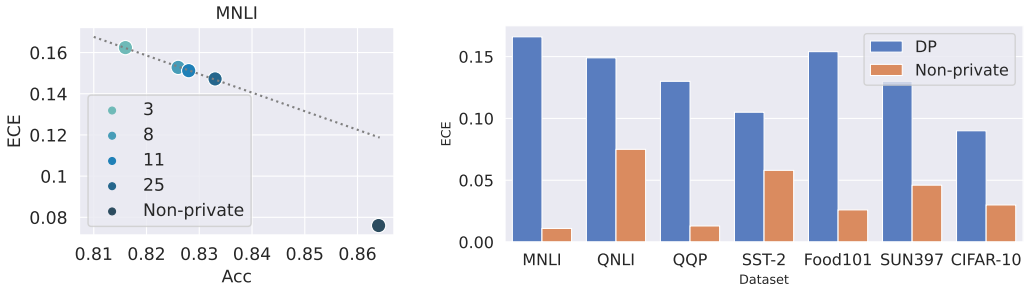
Ablation on per-example gradient clipping and noisy gradients. Differentially private SGD consists of two components: per-example gradient clipping and noise injection. We adjust the balance of clipping and gradient noise to understand which component contributes most to miscalibration.



(a) Synthetic results (b) MNLI results (c) Gradient clipping ablation

Figure 4: Per-example gradient clipping ($\epsilon = 8$) causes large ECE errors in (a) Logistic regression over non-separable 2D Gaussian data and (b) RoBERTa over MNLI. (c) Performing only gradient noising leads to high accuracy and low ECE.

In our synthetic two-dimensional Gaussian example (Fig. 4(a)), we see a clear trend where fixing the overall privacy guarantee (ϵ) and increasing the clipping norm from DP (0.1), DP (1) to DP (10) does not affect the accuracy, but substantially decreases the calibration error. The same setting on MNLI (Fig. 4(b)) also shows that increasing gradient norms slightly decreases ECE without impacting model accuracy. Finally, completely removing clipping and performing noisy gradient descent dramatically reduces ECE (and increases accuracy) in Fig. 4(c). Together, these results suggest that intensive clipping exacerbates miscalibration, even at the same privacy level.



(a) Privacy budget ϵ ablation (b) Comparison with non-private models with matched accuracies

Figure 5: (a) MNLI performance under varying **privacy budgets** ϵ. (b) **Controlling for accuracy** by early stopping non-private models to match the DP models does not substantially affect differences in ECE. The accuracy differences are within 1%.

Controlling for accuracy and regularization Accuracy and calibration are generally positively correlated (Minderer et al., 2021), which poses a question: does the miscalibration of DP models arise from lower accuracy in them? We find evidence against this possibility across two different experiments. First, we vary ϵ for DP in the MNLI experiment, which results in several models along a linear ECE-accuracy tradeoff (Fig. 5(a)). When compared to these private models, the non-private model has substantially lower ECE than would be expected by this tradeoff alone.

In a second experiment, we controlled the in-domain accuracy of non-private models to match their private counterparts by early-stopping the non-private models to within 1% of the DP model accuracy. Our results in Fig. 5(b) show that the ECE gap between private and non-private model persists even when controlling for accuracy.

More generally, we find that regularization methods such as early stopping impact the ECE-accuracy tradeoff qualitatively differently than DP-SGD. Our results in Tab. 5 show that most other regularizers such as early-stopping lead to an accuracy-ECE tradeoff, in which highly regularized models are less accurate but better calibrated. This is not the case with DP, where the model is both of lower accuracy and less calibrated relative to its non-private counterparts. This suggests that calibration errors in private and non-private settings may be caused by different reasons, and that DP-SGD-specific calibration methods could be even more effective.

Method	MNLI		Scitail		QNLI	
	Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
DP	0.826	0.1663	0.7707	0.222	0.5063	0.5090
Non-private	0.8642	0.0699	0.7853	0.1348	0.5050	0.4426
ℓ_2 (1e-4)	0.8664	0.0347	0.7876	0.0822	0.5058	0.4874
ℓ_2 (1e-3)	0.8672	0.0349	0.7891	**0.0816**	0.5056	**0.4862**
ℓ_2 (1e-2)	0.8620	**0.0326**	0.7845	0.0845	0.5059	0.4870
ℓ_2 (1e-1)	**0.8684**	0.1874	0.786	0.0835	0.5059	0.4872
dropout (0.1)	**0.8684**	0.1874	0.786	**0.0835**	0.5059	0.4872
dropout (0.2)	0.8601	**0.046**	0.7722	0.1076	0.5058	0.487
dropout (0.3)	0.8380	0.0629	0.7423	0.1523	0.5050	**0.486**
early stopping (2)	0.8423	**0.0288**	0.7806	**0.0662**	0.5050	**0.4818**
early stopping (4)	0.8572	0.0299	0.78	0.094	0.5058	0.486
early stopping (6)	0.8623	0.0355	0.7837	0.0811	0.5056	0.486
early stopping (8)	**0.8684**	0.1874	0.786	0.0835	0.5059	0.4872

Table 5: Comparison with non-private models trained using common **regularizers**, i.e. ℓ_2 (weight decay factor), dropout (probability) and early stopping (total training epochs). Models are trained on MNLI and evaluated over MNLI, Scitail and QNLI.

5 Concluding Remarks

In this work, we study the calibration of ML models trained with DP-SGD. We quantify the miscalibration of DP-SGD trained models and show they exist even using state-of-the-art pre-trained backbones. While the calibration errors are substantial and consistent, we show that adapting existing post-hoc calibration methods is highly effective for DP-SGD models. We believe it is an open question whether it is possible to leverage the generalization guarantees of DP-SGD to naturally obtain similarly well-calibrated models without the use of sample-splitting and recalibration.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *SIGSAC*, 2016. 1, 2, 3

Anastasios N Angelopoulos, Stephen Bates, Tijana Zrnic, and Michael I Jordan. Private prediction sets. *arXiv preprint arXiv:2102.06202*, 2021. 2

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *NeurIPS*, 2019. 1, 2

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, 2016. 1

Zhiqi Bu, Hua Wang, Qi Long, and Weijie J Su. On the convergence and calibration of deep learning with differential privacy. *arXiv preprint arXiv:2106.07830*, 2021. 1, 2, 3, 5, 13

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, 2021. 1

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022. 2, 5

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *EMNLP*, 2020. 2

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *RSS*, 2019. 3

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006. 3

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *STOC*, 2020. 1, 2

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *NeurIPS*, 2021. 3

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 4, 5, 13

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *PoPETs*, 2017. 1

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 2

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTail: A textual entailment dataset from science question answering. In *AAAI*, 2018. 5

Moritz Knolle, Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus R Makowski, Daniel Rueckert, and Georgios Kaissis. Differentially private training of neural networks with langevin dynamics for calibrated predictive uncertainty. *arXiv preprint arXiv:2107.04296*, 2021. 2, 3, 5, 13

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, pp. 2661–2671, 2019. 5

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosław Błasiok, and Preetum Nakkiran. What you see is what you get: Distributional generalization for algorithm design in deep learning. *arXiv preprint arXiv:2204.03230*, 2022. 1

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *ICLR*, 2022. 2, 5, 13

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

Rachel Luo, Shengjia Zhao, Jiaming Song, Jonathan Kuck, Stefano Ermon, and Silvio Savarese. Privacy preserving recalibration under domain shift. *arXiv preprint arXiv:2008.09643*, 2020. 2

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, 2019. 5

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *NeurIPS*, 2021. 2, 8

Ilya Mironov. Rényi differential privacy. In *CSF*, 2017. 3

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999. 4, 5

Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning ? In *AISTATS*, 2022. 1, 2

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *SP*, 2017. 1

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, 2013. 2

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *ICLR*, 2021. 2, 5

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. ICLR. 5

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *EMNLP*, 2020. 13

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021. 13

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, 2002. 4

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021. 1

A Privacy Analysis for Independent Releases With a Partition of Data

Our post-processing calibration setup requires splitting the original (private) training data into two disjoint splits where one of which is used solely for training and the other solely for post hoc recalibration. Given that both the training and post hoc recalibration algorithms are DP, it is natural to ask what is the overall privacy spending of the joint release. While one can essentially resort to any "off-the-shelf" privacy composition theorem, we note that in our setup the splits of data used in the two algorithms are disjoint, and thus a tighter characterization of privacy leakage is possible. The following is common knowledge, and we only include the proof for completeness.

Proposition A.1. *Let $M_1 : \mathcal{X}_1 \to \mathcal{Y}$ and $M_2 : \mathcal{X}_2 \times \mathcal{Y} \to \mathcal{Z}$ be (ϵ, δ)-DP algorithms consuming independent random bits operating on disjoint splits of the dataset. Then, the algorithm $M : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$ defined by*

$$M(X) = (y, z), \quad y = M_1(X_1), \quad z = M_2(X_2, y),$$

where (X_1, X_2) is a partition of X determined through some procedure independent on X, is also (ϵ, δ)-DP.

Proof. Let X and X' be neighboring datasets. Suppose that the first component in both partitions are the same, i.e., $X = (X_1, X_2)$, and $X' = (X_1, X_2')$, where X_2 and X_2' are neighboring. Then, M is (ϵ, δ)-DP directly follows from that M_2 is (ϵ, δ)-DP.

The more subtle case is when the second component in both partitions are the same. Specifically, suppose that $X = (X_1, X_2)$, and $X' = (X_1', X_2)$, where X_1 and X_1' are neighboring. Let R denote the random variable that controls only the randomness of M_2, i.e., conditioned a draw of $R = r$, M_2 is a deterministic function. With slight abuse of notation, we denote this deterministic function by $M_2(r)$. Let $O = \cup_{o_1 \in O_1} \{o_1\} \times O_2(o_1) \subset \mathcal{Y} \times \mathcal{Z}$ be a subset of the codomain. Define the following shorthand for the preimage of M_2 conditioned on $R = r$

$$M_2(r)^{-1}(X_2, S) = \{y \in \mathcal{Y} \mid M_2(r)(X_2, y) \in S\}.$$

Then, we have

$$\begin{aligned}
\Pr\left(M(X) \in O \mid R = r\right) &= \sum_{o_1 \in O_1} \Pr\left(M_1(X_1) = o_1\right) \Pr\left(M_2(X_2, o_1) \in O_2(o_1) \mid R = r\right) \\
&= \sum_{o_1 \in O_1} \Pr\left(M_1(X_1) = o_1\right) \mathbb{1}\left[M_2(r)(X_2, o_1) \in O_2(o_1)\right] \\
&= \sum_{o_1 \in O_1} \Pr\Big(M_1(X_1) = o_1,\ M_1(X_1) \in M_2(r)^{-1}(X_2, O_2(o_1))\Big) \\
&= \Pr\Big(M_1(X_1) \in \cup_{o_1 \in O_1} \Big(\{o_1\} \cap M_2(r)^{-1}(X_2, O_2(o_1))\Big)\Big) \\
&\leq e^{\epsilon} \Pr\Big(M_1(X_1') \in \cup_{o_1 \in O_1} \Big(\{o_1\} \cap M_2(r)^{-1}(X_2, O_2(o_1))\Big)\Big) + \delta \\
&= e^{\epsilon} \Pr\left(M(X') \in O \mid R = r\right) + \delta.
\end{aligned}$$

Since the above holds for all draws of R, we conclude that $\Pr\left(M(X) \in O\right) \leq e^{\epsilon} \Pr\left(M(X') \in O\right) + \delta$ for all neighboring X and X' which differ only in their first components. This concludes the proof. □

B Extended Experimental Details and Results

B.1 Settings for Synthetic Experiments

For **synthetic** experiments, we generate two-dimensional mixture Gaussian data of size $10k$. The distance between the centers of two class data shifts by a constant, which is set to be $2 * 1.5$. We use logistic regression to do the binary classification. We set the amount of data points from each class as 5k and batch size as $4k$. We include the results with different maximum gradient norm $C \in \{0.1, 0.5, 1\}$ of DP training.

Dataset	CIFAR-10	SUN397	Food101	MNLI	QNLI	QQP	SST-2
Learning rate	2e-3	1e-2	1e-4	5e-4	1e-3	5e-4	1e-3
Batch size	32	32	32	6,000	2,000	6,000	1,000
LR decay	False	False	False	True	True	True	True
Epochs	10	10	10	18	6	18	3
Weight decay	1e-4	1e-4	1e-4	0	0	0	0
Clipping norm	1.0	1.0	1.0	0.1	0.1	0.1	0.1
Privacy budget ϵ	3, 8	3, 8	3, 8	8	8	8	8
Validation ratio	0.1						
Noise scale	calculated numerically so that a DP budget of (ϵ, δ) is spent after E epochs						

Table 6: Default hyperparameter of DP finetuning over different datasets for reproducibility. Batch size is based on a unit batch size 20 with different amount of gradient accumulation steps. We use the validation ratio, the proportion of validation set, to split the training set for tuning recalibration methods.

B.2 Implementation Details

We use pre-trained checkpoints and trainers from Huggingface library (Wolf et al., 2020) for NLP experiments. We do linear probe for CV experiments using ResNet50 for CIFAR-10, ViT for SUN397 and Food101. We use the modified Opacus privacy engine (Yousefpour et al., 2021) from (Li et al., 2022), which computes per-example gradients for transformers. We compare DP training with popular regularizers used for finetuning like ℓ_2, dropout and early stopping over NLP datasets. ℓ_2 is the weight decay rate $\{1e-1, 1e-2, 1e-3, 1e-4\}$ during optimization. We apply dropout to both hidden and attention layers of transformers, which takes the value in $\{0.1, 0.2, 0.3, 0.4\}$. We do early stopping by setting the maximum amount of training epochs to be smaller, i.e. values in $\{2, 4, 6, 8\}$.

For recalibration training, we use a fixed amount of epochs without hyper-paramter tuning to avoid privacy leakage of validation sets. We initialize the temperature parameter in DP-TS as 1.0 and train 100 epochs for all the tasks except Food1001 (which uses 30 epochs) using DP-SGD with a 0.1 learning rate, 10 maximum gradient clipping norm, and a linearly decayed learning rate scheduler. We adapt multiclass extensions for Platt scaling by considering higher-dimensional parameters (Guo et al., 2017).

For baselines, we grid search the maximum norm bound $Z \in \{100, 500, 1000\}$ and epochs over $\{6, 8, 18\}$ for global clipping (Bu et al., 2021); we use pre-noise scale 0.046, temperature $\tau = 6.08$, exponential learning rate decay with a learning rate 0.005 and decay factor 0.028 as suggested by Knolle et al. (2021).

B.3 Additional Image Classification Results

In Tab. 7, we give additional results when we have smaller privacy budget $\epsilon = 3$. We see consistent results that DP fine-tuning gives poor calibration performance while DP-TS and/or DP-PS can recalibrate the classifiers effectively.

Category	Model	CIFAR-10		SUN397		Food101	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.7912	0.0916	0.6751	0.2806	0.7097	0.2464
	DP-SGLD	0.6953	0.1595	0.562	0.3295	0.6217	0.2834
	Global Clipping	0.7659	0.0782	0.6345	0.285	0.6853	0.2276
Recalibration	DP-PS	0.7823	**0.0109**	0.6694	0.2826	0.7084	0.0626
	DP-TS	0.7823	0.0217	0.6694	**0.0183**	0.7084	**0.0601**
Non-private	DP+Non-private-TS	0.7823	0.0218	0.6694	0.019	0.7084	0.0598
	Non-private	0.83	0.0794	0.7044	0.1062	0.8245	0.0349

Table 7: The image classification performance ($\epsilon = 3$) of different models before and after recalibration across datasets.

B.4 ADDITIONAL ABLATION STUDIES

Label noise injection. All of the datasets we consider have labels that are designed to be unambiguous, and the Bayes optimal predictor would produce a confidence histogram that is concentrated at 1.0. In this case, we might wonder whether the polarized confidence histograms observed in Fig. 3 are an artifact for datasets with unambiguous labels.

To understand this, we intentionally inject label noise into MNLI and study how this changes the behavior of DP-SGD and non-private learning algorithms. Specifically, we uniformly corrupt training labels - by selecting a uniform random class with probability $p \in \{0.6, 0.8\}$. We compare DP trained models and non-private models with 0.2 dropout regularization. The confidence histograms in Fig. 6 clearly demonstrate that differentially private models result in 100% confidence, *even when the Bayes optimal classifier can be at most 60% confident*. This shows that DP's miscalibration behavior that results in 100% confidence is not driven by a dataset's label distribution and this behavior is likely to be even worse on tasks with inherent label uncertainty.

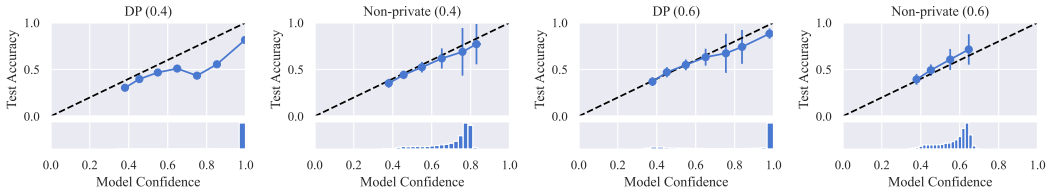


Figure 6: Reliability diagram for **label noise** settings with different models (corruption rates). For comparison, non-private models are included.