A Closer Look at the Calibration of Differentially Private Learners

Anonymous ACL submission

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 approaches for improving calibration with differential privacy only provide marginal improvements in calibration error while occasionally causing large degradations in accuracy. As a solution, we show that differentially private variants of post-processing calibration methods such as temperature scaling and Platt scaling are surprisingly effective and have negligible utility cost to the overall model. Across 7 tasks, temperature scaling and Platt scaling with DP-SGD result in an average 3.1-fold reduction in the in-domain expected calibration error and only incur at most a minor 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 in the form of training data leakage or 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, a commonly used DP algorithm to train machine learning (ML) models is DP-SGD (Abadi et al., 2016). The algorithm involves clipping per-example gradients and injecting noises into parameter updates during the optimization process.

Despite that DP-SGD can give strong privacy guarantees, prior works have identified that this privacy comes at a cost of other aspects of trustworthy ML, such as degrading accuracy and causing disparate impact (Bagdasaryan et al., 2019; Feldman, 2020; Sanyal et al., 2022). These tradeoffs pose a challenge for privacy-preserving ML, 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 tradeoff is significant as accessing model uncertainty is important for deploying models in safety-critical scenarios like healthcare and law where explainability (Cosmides and Tooby, 1996) and risk control (Van Calster et al., 2019) are needed in addition to privacy (Knolle et al., 2021).

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 promoting generalization (Dwork et al., 2015; Bassily et al., 2016; Kulynych et al., 2022). Moreover, training with modern pre-trained architectures show a strong positive correlation between calibration and classification error (Minderer et al., 2021), and using differentially private training based on pre-trained models are increasingly performant (Tramer and Boneh, 2021; Li et al., 2022b; De et al., 2022). However, we find that DP training has the surprising effect of consistently producing over-confident prediction scores in practice (Bu et al., 2021). We show an example of this phenomenon in a simple 2D logistic regression problem (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 transfers to more realistic settings.

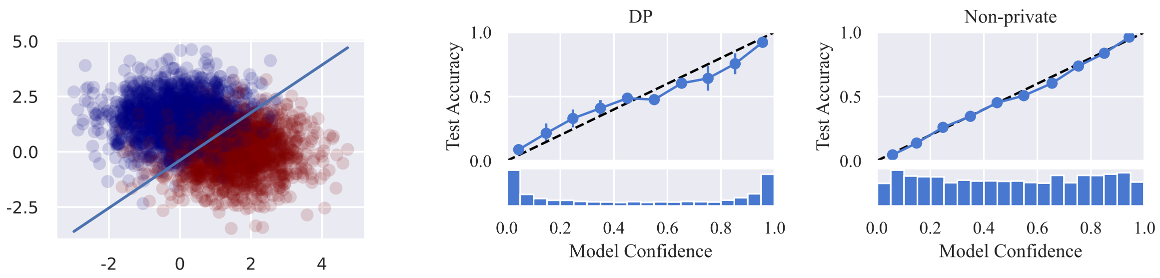
Our first contribution quantifies existing privacy-calibration tradeoffs for state-of-the-art models that leverage DP training and pre-trained backbones such as RoBERTa (Liu et al., 2019b) and vision transformers (ViT) (Dosovitskiy et al., 2020). Although there have been some studies of miscalibration for differentially private learning (Bu et al., 2021; Knolle et al., 2021), they focus on simple tasks (e.g., MNIST, SNLI) with relatively small neural networks trained from scratch. Our work shows that miscalibration problems persist even for state-of-the-art private models with accuracies approaching or matching their non-private counterparts. Through controlled experiments, we show that these calibration errors are unlikely solely due to the regularization effects of DP-SGD, and are more likely caused by the per-example gradient clipping operation in DP-SGD.

Our second contribution shows that the privacy-calibration 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, we show that DP-TS and DP-PS effectively prevent DP-trained models from being overconfident and give a 3.1-fold reduction in in-domain calibration error on average, substantially outperforming more complex interventions that have been claimed to improve calibration (Bu et al., 2021; Knolle et al., 2021).

2 Related Work

Differentially Private Deep Learning. DP-SGD (Song et al., 2013; Abadi et al., 2016) is a popular algorithm for training deep learning models with DP. Recent works have shown that fine-tuning high-quality pre-trained models with DP-SGD results in good downstream performance (Tramer and Boneh, 2021; Li et al., 2022b; De et al., 2022; Li et al., 2022a). 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; Tran et al., 2021; Sanyal et al., 2022; Esipova et al., 2022) (measured by the disparity in accuracies across groups). Our miscalibration findings are closely related to the above privacy-fairness tradeoff that has already received substantial attention. For example, per-example gradient clipping is shown to exacerbate accuracy disparity (Tran et al., 2021; Esipova et al., 2022). Some fairness notions also require calibrated predictions such as calibration over demographic groups (Pleiss et al., 2017; Liu et al., 2019a) or a rich class of structured "identifiable" subpopulations (Hébert-Johnson et al., 2018; Kim et al., 2019). Our work expands the 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 ML models. Recent works have observed that highly-accurate models that leverage pre-training are often well-calibrated (Hendrycks et al., 2019; Desai and 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 DP, 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) studied private calibration for out-of-domain settings, but did not study whether DP-SGD causes miscalibration in-domain. (Angelopoulos et al., 2021) modified split conformal prediction to be privacy-preserving, but they only studied vision models and their private models have substantial performance decrease compared to non-private ones. They also did not study the miscalibration of private models and the causes of the privacy-calibration tradeoff. (Knolle et al., 2021) studied 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 recalibration approach is consistently more effective. Closer to our work, the work by (Bu et al., 2021) identified that DP-SGD produces miscalibrated models on CIFAR-10, SNLI, and MNIST. As a solution, they suggested an alternative clipping scheme that empirically reduces the 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 transfer learning settings that are closer to the state-of-the-art setup in differentially private learning and find substan-



(a) Non-separable Gaussian Data

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

Figure 1: **DP-SGD gives rise to miscalibration for logistic regression.** (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 $b = (0, 1.5)$, and y is Rademacher distributed. (b) Reliability diagram and confidence histogram. 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**).

tially worse ECE gaps (e.g. they identify a 43% relative increase in ECE on CIFAR-10, while we find nearly 400% on Food101); we compare our simple recalibration procedure to their method and find that DP-TS is substantially more effective at reducing ECE.

Our main goal is to build classifiers that are both accurate and calibrated under differential privacy. We begin by defining core preliminary concepts.

2.1 Differential Privacy

Differential privacy is a formal privacy guarantee for a randomized algorithm which intuitively ensures that no adversary has a high probability of identifying whether a record was included in a dataset based on the output of the algorithm. Throughout our work, we will study models trained with approximate-DP / (ϵ, δ)-DP algorithms.

Definition 2.1. (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 $Y \subset \mathcal{Y}, \mathbb{P}(\mathcal{M}(X) \in Y) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{M}\left(X'\right) \in Y\right) + \delta$.

2.2 Differentially Private Stochastic Gradient Descent

The standard approach to train neural networks with DP is using 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 $\mathcal{B}_t$ is defined as

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

where η_t is the learning rate at step t, $\mathcal{L}\left(\theta^{(t)}\right)$ is the learning objective, $\operatorname{clip}_C\left(\nabla \mathcal{L}_i\left(\theta^{(t)}\right)\right)$ clips the gradient using $\operatorname{clip}_C\left(\nabla \mathcal{L}_i\left(\theta^{(t)}\right)\right) = \nabla \mathcal{L}_i\left(\theta^{(t)}\right) \cdot \min\left(1, C/\|\nabla \mathcal{L}_i\left(\theta^{(t)}\right)\|_2\right)$ and ξ is 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 as the noise multiplier returned by accounting and the expected batch size B. Each step of DP-SGD is approximate-DP, and the final model satisfies approximate-DP with privacy leakage parameters that can be computed with privacy loss composition theorems (Abadi et al., 2016; Mironov, 2017; Wang et al., 2019b; Dong et al., 2019; Gopi et al., 2021).

2.3 Calibration

A probabilistic forecast is said to be *calibrated* if the forecast has accuracy p on the set of all examples with confidence p. Specifically, given a multi-class classification problem where we want to predict a categorical variable Y based on the observation X, we say that a probabilistic classifier h_θ parameterized by θ over C classes satisfies *canonical calibration* if for each p in the simplex Δ^{C-1} and every label y, $P(Y = y \mid h_\theta(X) = p) = p_y$ holds.[1] Intuitively, a calibrated model should give predictions that can truthfully reflect the predictive uncertainty, e.g., among the samples to which a calibrated classifier gives a confidence 0.1 for class k, 10% of the samples actually belong to class k.

[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 (Guo et al., 2017). 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_θ is calibrated if

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

where p^* is the true predictive uncertainty. With the same definition of p^*, we will quantify the degree to which a classifier is calibrated through the expected calibration error (ECE), defined by

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

In practice, we estimate ECE by first partitioning the confidence scores into M bins $B_1, \ldots, B_M$ before calculating the empirical estimate of ECE as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}\left(B_m\right) - \mathrm{conf}\left(B_m\right)\right|, \tag{3}$$

where

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(y_i = \arg\max h_\theta(\mathbf{x}_i)), \tag{4}$$

$$\mathrm{conf}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} h_\theta(\mathbf{x}_i) \tag{5}$$

and $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ are a set of n i.i.d. samples that follow a distribution $P(X, Y)$. When appropriate, we will also study fine-grained miscalibration errors through the histogram of conf(B_m) (the confidence histogram) and plot acc(B_m) against conf(B_m) (the reliability diagram).

3 Experimental Results

We study three different experimental settings. We first consider **in-domain** evaluations, where we evaluate calibration errors on the same domain that they are trained on. Results show that using pre-trained models does 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 recalibration methods carry over to the out-of-domain setting. Finally, we perform careful **ablations** to isolate and understand the causes of in-domain miscalibration. In each case, we will show that DP-SGD leads to high miscalibration, and DP recalibration substantially reduces calibration errors.

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 fine-tune RoBERTa-base using the procedure in (Li et al., 2022b), and for vision datasets, we perform linear probe of ViT and ResNet-50 features, following (Tramer and Boneh, 2021).

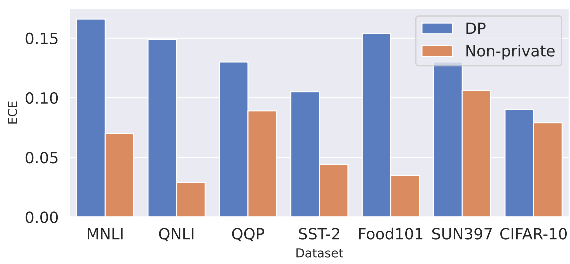


Figure 2: DP trained models display consistently higher ECE than their non-private counterparts.

Datasets. Following prior work (Li et al., 2022b), we train on MNLI, QNLI, QQP, SST-2 (Wang et al., 2019a) for the text classification 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., 2019a).[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 train 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). The last two methods are included to evaluate our simple recalibration approaches against existing methods which are reported to improve calibration.

For our recalibration methods, we run the private recalibration method over the in-domain recalibration set X_{recal} in Sec. 3.1 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 com-

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

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

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

Category	Model	MNLI		QNLI		QQP		SST-2	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.8281	0.166	0.8503	0.149	0.8685	0.13	0.8922	0.105
	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

bines differentially private model training with non-private temperature scaling (DP+NON-PRIVATE-TS) as a way to quantify privacy costs in the post-hoc recalibration step. Further implementation details and default hyper-parameters for DP training are in Tab. 6 in Appendix B.

3.1 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 and Boneh, 2021; Li et al., 2022b; De et al., 2022).

However, we find that these same models have substantially higher calibration errors. For example, the linear probe for Food101 in Fig. 2 has private accuracy within 7% of the non-private counterpart, but the ECE is more than 4× that of the non-private counterpart. In the language case, we see similar results on QNLI with a ~6% decrease in accuracy but a ~4.3× 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 well consistently over all datasets and on both modalities with marginal accuracy drops (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 can have very close or even 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 confidences 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, after recalibration, we see almost perfect agreement between the model confidences and actual accuracies.

3.2 Out-of-domain Calibration

We complement our in-domain experiments with out-of-domain evaluations. To do this, we eval-

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

Dataset	Category	Model	Hans		Scitail		RTE		WNLI	
			Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
MNLI	Baseline	DP	0.5195	0.4786	0.7761	0.2172	0.7437	0.2541	0.4507	0.5492
		DP-SGLD	0.4996	0.4995	0.7515	0.233	0.6498	0.3169	0.4507	0.5491
		Global Clipping	0.5221	0.4747	0.7845	0.2051	0.7076	0.2737	0.4366	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	0.5046	0.4932	0.729	0.2666	0.5657	0.4407	0.4724	0.5215
		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

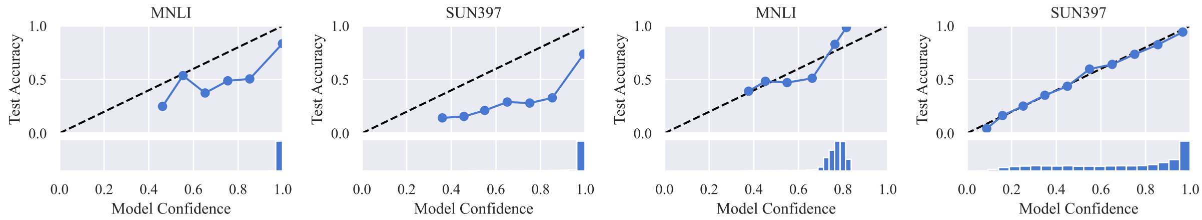


Figure 3: Reliability diagram and confidence histogram before (**Left**) and after (**Right**) recalibration using DP-TS. Recalibration parameters are learned on the validation set X_{recal} of MNLI and SUN397.

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

Dataset	Category	Model	MRPC	
			Accuracy	**ECE**
QQP	Baseline	DP	0.7475	0.252
		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

uate the zero-shot transfer performance of models trained over MNLI, QNLI (Tab. 3) and QQP (Tab. 8).

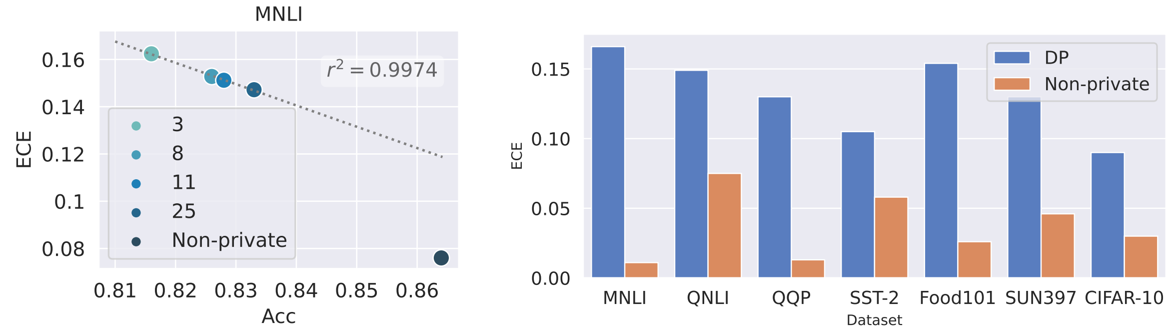
Our findings are consistent with the in-domain evaluations. Differentially private training generally results in high ECE, while DP-TS and DP-PS generally improve calibration. The gaps out-of-domain are substantially smaller than the in-domain case, as all methods are of low accuracy and miscalibrated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

3.3 Analyses and Ablation Studies

Finally, we carefully study two questions to better understand the 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 noise injection. DP-SGD involves per-example gradient clipping and noise injection. To better understand which component contributes more to miscalibration, we perform experiments to isolate the effect of each individual component.

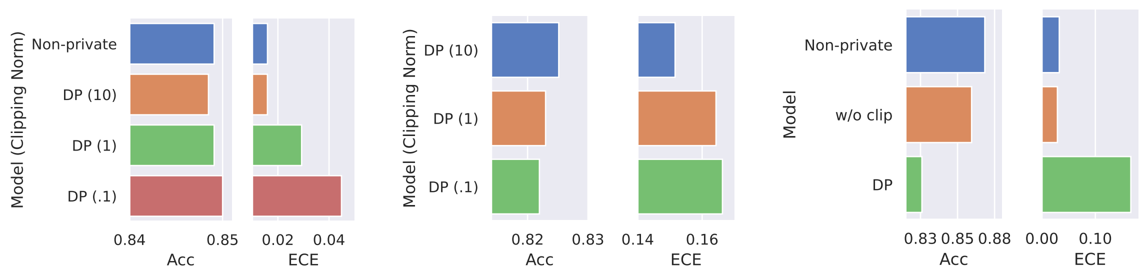
On 2D synthetic data (example given in Fig. 1), Fig. 5(a) shows that fixing the overall privacy guarantee (ϵ) and increasing the clipping threshold from DP (0.1) to DP (1) and further to DP (10) affect the accuracy only marginally but substantially improve calibration. Repeating this ablation with RoBERTa fine-tuning on MNLI (Fig. 5(b)) confirms that increasing the clipping threshold (slightly) decreases ECE but does not substantially impact model accuracy. Finally, Fig. 5(c) shows that completely removing clipping and training with only noisy gradient descent dramatically reduces ECE (and increases accuracy). These results suggest that intensive clipping exacerbates miscalibration (even



(a) Privacy budget ϵ ablation

(b) Comparison with non-private models with matched accuracies

Figure 4: (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%.



(a) 2D synthetic data

(b) MNLI

(c) Gradient clipping ablation

Figure 5: Per-example gradient clipping ($\epsilon = 8$) causes large ECE errors in (a) logistic regression on non-separable 2D synthetic data, and (b) fine-tuning RoBERTa on MNLI. (c) Performing only gradient noising leads to high accuracy and low ECE.

under a fixed privacy guarantee).

Controlling for accuracy and regularization. Accuracy and calibration are generally positively correlated (Minderer et al., 2021; Carrell et al., 2022). This poses a question: Does the miscalibration of DP models arise due to their suboptimal accuracy? We find evidence against this in two different experiments.

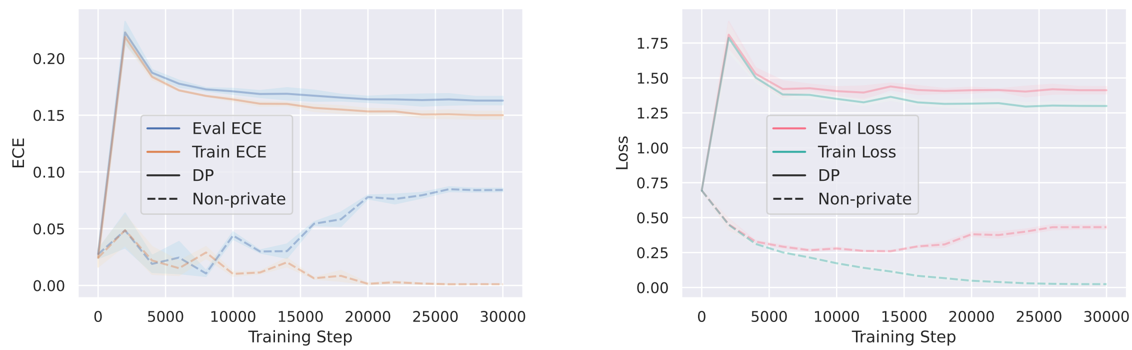
In the first experiment, we vary ϵ for fine-tuning RoBERTa with DP on MNLI. This results in several models situated on a linear ECE-accuracy tradeoff curve (Fig. 4(a)). Intuitively, extrapolating this curve helps us identify the anticipated ECE for a DP trained model with a given accuracy. Fig. 4(a) shows that when compared to these private models, the non-private model has substantially lower ECE than would be expected by extrapolating this tradeoff alone. This suggests that private learning experiences a qualitatively different ECE-accuracy tradeoff than standard learning.

In the 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 be within 1% of the DP model accuracy. Fig. 4(b) shows that the ECE gap between the private and non-private models 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 for DP training, where the resulting models are both of lower accuracy and less calibrated relative to their non-private counterparts. These findings suggest that calibration errors in private and non-private settings may be caused by different reasons - the miscalibration of private models may not be due to the regularization effects of DP-SGD.

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.

Method	MNLI		Scitail		QNLI	
	Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
DP	0.8281	0.166	0.7761	0.2172	0.5058	0.4942
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



(a) ECE (b) Loss

Figure 6: **DP-SGD training** ($\epsilon = 8$) **makes train and eval ECE close but both of them are large**. The training dynamics of (a) ECE and (b) Loss on both QNLI training and evaluation sets.

DP training leads to similarly high train and test ECE. Learning algorithms which satisfy tight DP guarantees are known to generalize well, meaning that the train (empirical) and test (population) losses of a DP trained model should be similar (Dwork et al., 2015; Bassily et al., 2016). In a controlled experiment, we fine-tune RoBERTa on QNLI with DP-SGD ($\epsilon = 8$) and observe that the train-test gaps for both ECE and loss are smaller for DP models than the non-private ones (Fig. 6). Yet, for DP trained models, both the train and test ECEs are high compared to the non-private model. Interestingly, these observations with DP trained models are very different from what's seen in miscalibration analyses of non-private models. For instance, (Carrell et al., 2022) showed that non-private models tend to be calibrated on the training set but can be miscalibrated on the test set due to overfitting (large *calibration generalization gap*). Our results show that DP trained models have a small calibration generalization gap, but are miscalibrated on both the training and test sets.

4 Discussion and 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 verify that 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. 2016. Deep learning with differential privacy. In *SIGSAC*.

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

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

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

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

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

Annabelle Carrell, Neil Mallinar, James Lucas, and Preetum Nakkiran. 2022. The calibration generalization gap. *arXiv preprint arXiv:2210.01964*.

Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*.

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

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

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

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

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. 2015. Preserving statistical validity in adaptive data analysis. In *STOC*.

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

Maria S Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C Cresswell. 2022. Disparate impact in differential privacy from gradient misalignment. *arXiv preprint arXiv:2206.07737*.

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

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

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

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

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. In *ICML*.

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

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

Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *AIES*.

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

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

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

Xuechen Li, Daogao Liu, Tatsunori Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin Tat Lee, and Abhradeep Guha Thakurta. 2022a. When does differentially private learning not suffer in high dimensions? *NeurIPS*.

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

Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019a. The implicit fairness criterion of unconstrained learning. In *ICML*.

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

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

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

Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30.

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

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

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

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *NeurIPS*.

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

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

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

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

Cuong Tran, My Dinh, and Ferdinando Fioretto. 2021. Differentially private empirical risk minimization under the fairness lens. *NeurIPS*.

Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. 2019. Calibration: the achilles heel of predictive analytics. *BMC medicine*.

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

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. 2019b. Subsampled rényi differential privacy and analytical moments accountant. In *AISTATS*.

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

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

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

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. Based on parallel composition (McSherry, 2009), we prove the following proposition.

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 is 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 is 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

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

Since the above holds for all draws of R, we conclude that $\Pr\left(M(X) \in O\right) \le e^{\epsilon} \Pr\left(M(X') \in O\right) + \delta$ 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.

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.

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						

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., 2022b), 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 $\{2, 4, 6, 8\}$. The default hyper-parameters for ℓ_2, dropout, early stopping are $1e-1$, 0.1, 8 respectively so some of the results in Tab.5 are reused.

For recalibration training, we use a fixed amount of epochs without hyper-parameter 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 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 a 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.

B.4 Additional Text Classification Results

We include additional text classification results (Tab. 8 and Tab. 9) and see a consistent trend that DP training leads to higher ECE than the non-private ones. DP-TS and DP-PS can give reduction on ECE even without further training on target domains.

B.5 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.

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

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

Dataset	Category	Model	MRPC	
			Accuracy	**ECE**
QQP	Baseline	DP	0.7475	0.252
		DP-SGLD	0.6936	0.3016
		Global Clipping	0.7475	0.252
	Recalibration	DP-PS	0.7426	**0.1317**
		DP-TS	0.7426	0.1765
	Non-private	DP+Non-private-TS	0.7426	0.1765
		Non-private	0.7255	0.2671

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

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-SGD trained models and non-private models with 0.2 dropout regularization. The confidence histograms in Fig. 7 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-SGD trained model's miscalibration behavior that results in near 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.

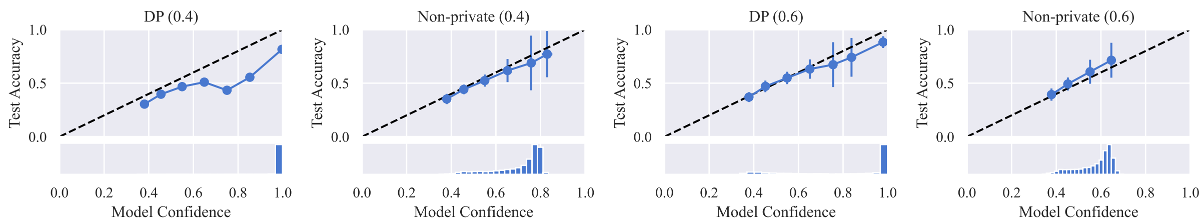


Figure 7: Reliability diagram and confidence histogram for **label noise** settings with different models (corruption rates) trained on MNLI. For comparison, non-private models are included.

C Remarks on Correlations between Accuracy and Calibration

In general, the correlations between accuracy and calibration are not clearly understood even for non-private learners as many factors can impact calibration such as architecture, regularization, optimization, data distribution, overparameterization, etc. Below we include some notable empirical findings. Convolutional networks like ResNets and DenseNets can be miscalibrated (Guo et al., 2017). However, (Minderer et al., 2021) show that modern models like ViT (Dosovitskiy et al., 2020) are better calibrated compared to past models; modern neural networks tend to have a strong positive correlation between calibration and classification error; model architectures matter greatly in calibration properties. Using pre-training can

Dataset	Category	Model	Hans		Scitail		RTE		WNLI	
			Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
QNLI	Baseline	DP	0.5002	0.5	0.7377	0.7351	0.5632	0.5546	0.5352	0.4663
		DP-SGLD	0.5	0.4986	0.7209	0.7148	0.5668	0.5557	0.5775	0.4235
		Global Clipping	0.5025	0.5021	0.7293	0.7272	0.5198	0.5219	0.5211	0.4778
	Recalibration	DP-PS	0.5002	**0.3452**	0.7377	**0.5793**	0.5632	**0.3924**	0.5352	**0.3073**
		DP-TS	0.5002	0.4058	0.7377	0.629	0.5632	0.4426	0.5352	0.3653
	Non-private	DP+Non-private-TS	0.5002	0.4058	0.7377	0.629	0.5632	0.4426	0.5352	0.3653
		Non-private	0.538	0.273	0.7454	0.5646	0.5199	0.3433	0.451	0.4035

Table 9: Additional **zero-shot transfer** NLI performance ($\epsilon = 8$) from QNLI to multiple OOD test datasets.

improve model uncertainty and calibration (Hendrycks et al., 2019; Desai and Durrett, 2020; Minderer et al., 2021; Kadavath et al., 2022). Regularizations like gradient noise injection can promote stability and distributional generalization so good calibration over the training set can transfer to the test set (Kulynych et al., 2022). (Carrell et al., 2022) empirically shows that popular models with small generalization gaps will have small test calibration errors.

Realizing the above observations, it is possible that the per-example gradient clipping and gradient noise injection in DP-SGD can contribute to both accuracy and calibration in different ways. Therefore, we carefully control the accuracy and regularization when conducting analyses and drawing conclusions (Tab. 5, Fig. 4(a) and 4(b), Fig. 7). However, even with the confounding controls above, DP-SGD trained models are still miscalibrated. In other words, the reason for the finding that private learners are much more miscalibrated than non-private counterparts is less likely to be the unambiguous labels in datasets, accuracy discrepancy or regularization effects of DP-SGD but more likely to be the per-example gradient clipping operation.

a classifier with high accuracy does not necessarily have good calibration. For example, a highly accurate but miscalibrated classifier can always output polarized confidence scores so the top-class confidence is always above 0.9. This is corroborated both qualitatively (Fig. 3) and quantitatively (Tab. 1 and Tab.2) in our experiments.