LLM-based Privacy Data Augmentation Guided by Knowledge Distillation with a Distribution Tutor for Medical Text Classification

Yiping Song[1,*], **Juhua Zhang**[2,*], **Zhiliang Tian**[2,†], **Yuxin Yang**[2], **Minlie Huang**[3], **Dongsheng Li**[2]
[1]College of Science, National University of Defense Technology, Hunan, China
[2]College of Computer, National University of Defense Technology, Hunan, China
[3]The CoAI Group, DCST, BNRist, Tsinghua University, Beijing 100084, China
{songyiping, zhangjuhua23, tianzhiliang,
yangyuxin21a, dsli}@nudt.edu.cn, aihuang@tsinghua.edu.cn

Abstract

As sufficient data are not always publically accessible for model training, researchers exploit limited data with advanced learning algorithms or expand the dataset via data augmentation (DA). Conducting DA in private domain requires private protection approaches (i.e. anonymization and perturbation), but those methods cannot provide protection guarantees. Differential privacy (DP) learning methods theoretically bound the protection but are not skilled at generating pseudo text samples with large models. In this paper, we transfer DP-based pseudo sample generation task to DP-based generated samples discrimination task, where we propose a DP-based DA method[1] with a LLM and a DP-based discriminator for text classification on private domains. We construct a knowledge distillation model as the DP-based discriminator: teacher models, accessing private data, teaches students how to select private samples with calibrated noise to achieve DP. To constrain the distribution of DA's generation, we propose a DP-based tutor that models the noised private distribution and controls samples' generation with a low privacy cost. We theoretically analyze our model's privacy protection and empirically verify our model.

1 Introduction

The ability of deep text classification models mainly derives from large-scale training data and recent large language models (LLMs) particularly benefit from big data. Many domains (e.g. medicine (Qing et al., 2019; Li et al., 2021)) have only limited amount of public data and large private data. It is risky to release private data to model training since models probably memorize details in those data and unintentionally output their sensitive information (Carlini et al., 2019a, 2021).

Researchers fully exploit limited public data and avoid accessing private data to ensure data security (Fei-Fei et al., 2006), which achieves learning on small data via meta-learning (Finn et al., 2017) or active learning (Konyushkova et al., 2017). Further, some researchers expand the public dataset with data augmentation (DA) (Wei and Zou, 2019); another promising DA strategy is to synthesize private samples conditioned on privacy data while ensuring the sensitive information in private data are well-protected (Yue et al., 2022).

Synthesizing private text requires to capture original private data distributions by accessing the data. Accessing them should be under the guarantee of privacy protections and thus we need privacy protection on text generation. Researchers attempt anonymization methods to mask selected sensitive text span (Carlini et al., 2019b; Shi et al., 2021), randomly noise the input tokens (Feyisetan et al., 2020), or apply regularization to avoid overfitting training data (Carlini et al., 2019b). However, the masking, nosing, or regularization mechanism hardly ensures (almost) all personal information is well-protected. Differential privacy (DP) (Dwork et al., 2006) provides provable guarantees against the identification of individual information in datasets. Deep generative models with DP ensure the existence of a specific sample (with private information) cannot be detected (Li et al., 2022).

Noisy-SGD (Song et al., 2013; Abadi et al., 2016) is a practical DP algorithm for deep learning models, including text generation models (Kerrigan et al., 2020), which adds calibrated noise on model gradients to satisfy DP. However, the advantage of NoisySGD is weaken as the model becomes larger (Bassily et al., 2014; Yu et al., 2020) and NoisySGD requires a per-example gradient clip resulting in non-convergence and system overheads (Zhu et al., 2020; Bu et al., 2021). Another categories of DP learning methods, PATE (Papernot et al., 2017), acquires private information from

[*] Equal Contribution.
[†] Corresponding Author.
[1] Code is available at https://anonymous.4open.science/r/DP_DA-1BF7

teacher models learned on private data to a student and adds noises on teachers' outputs. PATE's privacy cost comes from teachers' outputs instead of the whole model's gradients in Noisy-SGD (Li et al., 2022). Hence, PATE's required noise does not scale with model size and PATE is promising of working on large models (e.g. BERT (Devlin et al., 2019), Llama (Touvron et al., 2023), and GPT-4 (Achiam et al., 2023)).

DP with PATE on text generation still suffers from a sequential multiple tasks with large candidate space, which extremely (linearly) increases the noise scale (Tian et al., 2022). We argue that we can transfer a DP-based generation task to a DP-based discrimination task to avoid complexities in DP text generation models. Particularly, we employ LLMs' generation ability $P_{LLM}(x)$ to generate public samples x and construct a DP-based discriminator $P_{discri}(\cdot|x)$ to select synthesized samples x fits for private domain c, synthesizing private samples via $P_{syn}(x|\cdot) = P_{discri}(\cdot|x) * P_{LLM}(x)$.

In this paper, we propose a DP-based data augmentation (DA) paradigm with a LLM and a DP-based discriminator to generate samples for private text classification, where the discriminator selects LLM-generated samples likely to belong to private domain as our synthesized samples. Specifically, the discriminator achieves DP via knowledge distillation (KD), where multiple teacher models access disjoint and unique private sets to learn private discriminators; a student learns from noised aggregated teacher outputs to achieve DP. To control the distribution in DA's generated samples, we propose a DP-based distribution tutor that captures the distribution of private data. In DP, querying privacy is expensive (student querying teacher causes a certain privacy cost), the tutor carries less sensitive information than teachers and helps the student with a low privacy cost (See §4). We further provide theoretical analyses and empirical results to verify our methods.

Our contributions are as follows: (1) We construct a DP-based DA with LLMs that synthetizes (almost infinite) samples while bounding the privacy leakage. (2) We propose a DP-based tutor for teacher-student frameworks to teach some less sensitive data with a low privacy cost. (3) We excel strong DP-based baselines on text classification in private domains.

2 Related Work

2.1 Privacy Protection in Text Classification

Initial privacy protection techniques predominantly focused on data anonymization (Maeda et al., 2016; Suzuki et al., 2018), For example, de-identification techniques (Garfinkel et al., 2015) such as removing, replacing, or encrypting sensitive information in data can reduce the risk of privacy leaks. Data perturbation techniques (Johnson and Shmatikov, 2013) protect user privacy by incorporating random noise into the data. Nevertheless, straightforward data anonymization measures may difficult to effectively deal with privacy leakage challenges (Rocher et al., 2019). Presently, federated learning (McMahan et al., 2017; Deng et al., 2022) and DP (Dwork et al., 2006) emerge as the two principal methodologies in the domain of privacy protection. Federated learning strategies can prevent privacy leaks caused by untrustworthy servers. DP aims to prevent attackers from extracting sensitive information from the training dataset (Carlini et al., 2021), offering a quantifiable privacy protection mechanism. With its robust theoretical foundation and broad applicability(Dwork et al., 2014a), differential privacy is widely acknowledged as the standard practice in the field of privacy protection. Independently and concurrently with our work, (Wu et al., 2023) and (Duan et al., 2023) studied ICL with DP guarantees for text classification tasks. Our proposed method predominantly embraces the privacy protection principles of differential privacy.

2.2 Data Augmentation (DA) in Text Classification

Various NLP data augmentation (DA) techniques have been developed, such as Back-Translation (Kobayashi, 2018), EDA (Wei and Zou, 2019), and AEDA (Karimi et al., 2021). These methods primarily focus on modifying the original input, which limits the diversity of the generated samples. In response, (Szegedy et al., 2016) initially explore a interpolation-based methods (i.e., mixup) in computer vision. Subsequently, (Guo et al., 2019) combine the mixup technique with CNNs and LSTMs for text applications.There are also many studies that choose different strategies to improve the mixup technique (Chen et al., 2020; Zhang et al., 2020).

Moreover, some researchers use pre-trained language models (PLMs) for data for data augmentation. (Kumar et al., 2020) provide a straightforward and effective method for conditional PLM by prepending class labels to text sequences. (Hu et al., 2019) utilize reinforcement learning with a conditional language model that performs by appending the correct label to the input sequence during training. Further, an increasing number of scholars have started to utilize adversarial learning methods to generate augmented samples, such as BERT-Attack (Li et al., 2020), G-DAUG^C (Yang et al., 2020).

To reduce the negative impact of low-quality augmentation samples on model performance, some research focus on sample selection. For example, (Cao et al., 2021) propose UAST framework to quantify model uncertainty for selecting pseudo-labeled samples. (Lin et al., 2023) focus more on the combination of sample selection and data enhancement strategies, and introduce a self-training selection framework to select high-quality samples from the data augmentation. Different from the above methods, we aim to achieve DA to synthetic private data while ensuring the private information from the private dataset.

2.3 DP for Deep Learning Models

Some researchers employ DP to protect the privacy of empirical risk minimization classifiers (Chaudhuri et al., 2011) and SVM (Rubinstein et al., 2009). Following (Song et al., 2013), NoisySGD introduces noise into gradients to achieve DP for deep learning models, including DP-SGD (Song et al., 2013; Bassily et al., 2014; Bu et al., 2023) and DP-Adam (Abadi et al., 2016; Kingma and Ba, 2014). The use of DP-SGD for large-scale pre-training of BERT has been shown to achieve comparable masked language modeling performance to non-private BERT (Anil et al., 2021), but with a privacy budget is 100 or higher. Recent studies have demonstrated that even under more stringent privacy constraints, generative and discriminative language models can achieve high performance across various tasks by appropriately selecting hyperparameters and fine-tuning objectives aligned with the pre-training process (Li et al., 2022). Additionally, (Li et al., 2022) apply ghost clipping to pre-trained language models using NoisySGD, reducing memory usage. (He et al., 2022) leverage group clipping with adaptive clipping thresholds, privately fine-tuning GPT-3 with 1.75 trillion parameters. PATE (Papernot et al., 2017) is another type of DP learning algorithm, transferring knowledge from teacher models trained on private sets with noises to a student model. The privacy cost of PATE arises from knowledge distillation rather than the gradient of the entire model. With this advantage, PATE has enormous potential in adapting to large models. Moreover, PATE is designed for classification tasks and is suitable for our goal of training a DP-based discriminator.

3 Methods

3.1 Overview

Our model consists of four parts (Fig. 1).

LLM-based Public Generator (§3.3) synthetic public textual input for the specific label on the text classification task.

DP-based Discriminator with Knowledge Distillation (§3.4) learns to discriminate whether a sample is (likely) from public domain or private domain, which satisfy DP privacy guarantee. It filters the samples from the generator (§3.3) to obtain new samples for privacy domain.

Label Distribution Tutor (§3.5) leads the generated pseudo samples to follow the distribution of private data under DP guarantee, which also filter the generator §3.3's output.

Private Data Augmentation (§3.6) uses the generator to obtain candidate samples and uses the discriminator (§3.4) and the tutor (§3.5) to filter the candidates to get data augmentation samples.

3.2 Task Definition

Given a private text classification corpora with sensitive information, our task aims to train a text classifier with a certain level of theoretical privacy guarantee, which protects all training samples with its label $< x, c >$ from being detected (protecting the existence of specific training sample from being identified by any detectors via any detecting methods).

Note that our task allows the model using data augmentation (DA) method to generate pseudo private samples for training. However, the data augmentation model also ought to ensure the privacy guarantee mentioned above.

3.3 LLM-based Public Generator

We employ a LLM (i.e. GPT-3.5) to generate input texts for each output label (on classification task). Even if the GPT-3.5 is trained on the public domain, we design a prompt text to induce the LLM to attempt to generate private samples, where

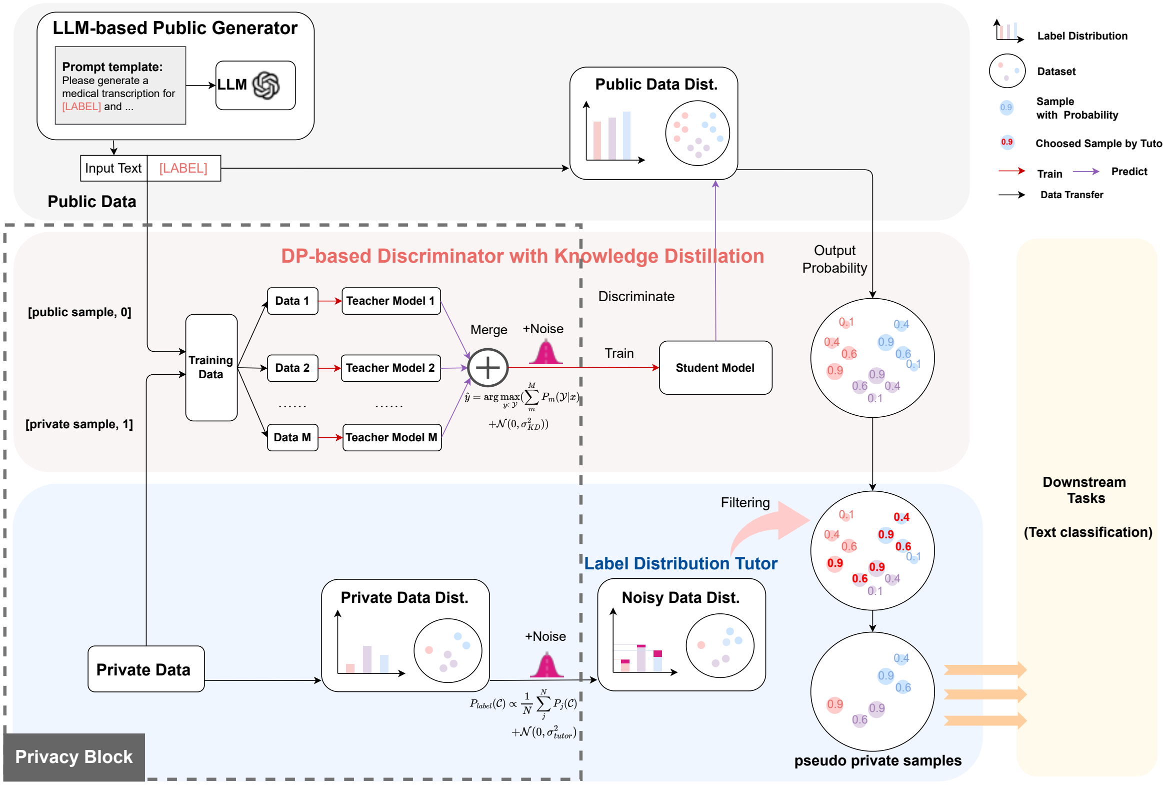

Figure 1: Overview of our framework. It mainly contains three components, LLM-based Public Generator (grey block) generates public data. DP-based Discriminator with Knowledge Distillation (pink block) discriminates public data and obtains a probability similar to private data. The Label Distribution Tutor (blue block) selects a subset with the highest probabilities of samples matching the noise label distribution. The gray dotted box is the privacy block.

the prompt follows this template: “You are a professional medical transcriber. Please generate a medical transcription for [LABEL] and do not reveal the patient’s name. The text length of medical transcription is approximately 400 words and at least 200 words.” In the above template, “[LABEL]” denotes the label c of classifications (i.e. medical specialty). In this way, LLM generates the input text x with the label c to compose a complete pseudo sample $< x, c >$.

3.4 DP-based Discriminator with Knowledge Distillation

We propose a DP-based discriminator to check if the pseudo samples fit for the private distribution thus are capable of acting as private samples. Inspired by (Papernot et al., 2017), we construct a teacher-student framework with knowledge distillation. This model consists of multiple teacher models and a student model. Teachers are allowed to access the private data and the student can only access the noised teachers’ outputs.

Teacher Models. We train multiple teacher models on multiple disjoint datasets, where the private samples act as the positive sample and the public samples (generated by §3.3) as the negative sample. The teachers learn to judge whether the samples from private set or public set. All teachers follow the structure and the initial parameters of a pretrained model.

To satisfy DP, we carefully process the teachers’ data and train teachers with two strategies. First, the private training data need to contain only unique samples (private sample duplicates should be removed). The reason is that DP prevents the unknown detectors from identification on each occurrence (Dwork et al., 2006). Duplicated samples with N occurrences increase the private loss and thus decrease the bound of privacy protection (in terms of ε) (Dwork et al., 2014a). To maintain the bound, the noise scale has to increase N times (Kerrigan et al., 2020) (or $\log(N)$ times s.t. advanced DP (Li et al., 2022)) which extremely harms for the performance.

Second, we equally divide the shuffled private data into M disjoint sets for M teachers and train each teacher on each set separately. In this way, the existence of any specific sample affects only one teacher's output, which bounds the sensitivity of the teachers' noisy distribution (Papernot et al., 2017; Boenisch et al., 2023) (See detailed analyses in §4). The equally divide (in terms of the sample number) and shuffled assignment ensure the balance among all teachers and the training performance.

Student Model. The student model is a discriminator cannot access raw private data and learn from teachers with those steps: (1) `Merging`. For a given sample x, we merge teachers' predictions $P_m(\mathcal{Y}|x)$, where $\mathcal{Y} = \{0, 1\}$ indicates whether the sample x derives from private or public set. (2) `Noising`. According to DP theory (Dwork et al., 2014b), we add calibrated noise on the teacher's predictions to obtain $\hat{y}$ as Eq. 1, denoting teachers' estimated judgment on x.

$$\begin{aligned}\hat{y} &= f_{tea\_agg}(\mathcal{Y}|x) \qquad (1)\\ &= \arg\max_{y\in\mathcal{Y}}(\sum_{m}^{M} P_m(\mathcal{Y}|x) + \mathcal{N}(0, \sigma^2_{KD}))\end{aligned}$$

, where the first term is the actual teacher aggregated outputs and M indicates the teacher number; the second term represents the calibrated Gaussian noise (Bu et al., 2020) and $\sigma_{teacher}$ controls the degree of adding noise (i.e. degree of privacy protection). (3) `Teaching`. We employ the noisy outputs $\hat{Y}$ to act as pseudo labels, which teaches the student how to discriminate the privacy samples. For each sample x, the teaching follows Cross-Entropy loss as Eq. 2.

$$\begin{aligned}\mathcal{L}_{KD} &= \mathrm{CE}(f_{tea\_agg}(\mathcal{C}|x), P_{student}(\mathcal{C}|x)) \qquad (2)\\ &= -\sum_{m}^{M} \mathbb{I}(c_m = \hat{c}) \log P_{student}(c_m|x)\end{aligned}$$

, where $\mathbb{I}$ is a indicator function. The above mechanism ensures the student satisfies DP as (Papernot et al., 2017) (Analyses in §4).

3.5 Label Distribution Tutor

We propose a tutor based on the above teacher-student framework (§3.4), which avoids directly access the pure private samples but accesses only a small amount of privacy information so as to maintain the bound of protections. The tutor aims to carry the label distribution of the private samples, as the label distribution is critical in generating augmented data.

The tutor follows a statistic way to collect the noised label distribution of all private samples $P_{label}(\mathcal{C})$ with a very small privacy cost. The tutor first obtains the actual label distribution by counting the total N training samples as the first term of Eq. 3. Then, it imports calibrated Gaussian noise on the label distribution as the second term of Eq. 3 [2]. The above mechanism ensures the tutor satisfy DP (Analyses in §4).

$$P_{label}(\mathcal{C}) \propto \frac{1}{N}\sum_{j}^{N} P_j(\mathcal{C}) + \mathcal{N}(0, \sigma^2_{tutor}). \qquad (3)$$

3.6 Private Data Augmentation

We propose a novel data augmentation (DA) framework to generate samples for private domain while protecting the privacy. The framework involves: (1) a public LLM-based generator to obtain a plenty of candidate samples (§3.4). (2) a DP-based discriminator to select LLM's generated samples similar to private samples, and (3) a tutor to filter generated samples to ensure the label distributions fitting for the private data.

The idea is (1) taking advantage of strong generation ability of LLMs to obtain high quality data. (2) accessing private data causes privacy cost but accessing privacy through the student (i.e. discriminator) and tutor satisfying DP would not brings in additional loss. Hence, we can "infinitely" call LLMs, student, and tutor to achieve DA while bounding the privacy protection.

4 Privacy Analyses of our Method

Lemma 4 Analytical Gaussian mechanism. (Balle and Wang, 2018) For a query $h : \mathcal{X}^n \rightarrow \mathcal{Y}^d$ over a dataset $\mathcal{D}$, the randomized algorithm outputting $h(\mathcal{D}) + Z \;\; s.t. \;\; Z \sim \mathcal{N}(0, \sigma^2 I_d)$ satisfies $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geq 0$ and $\delta(\varepsilon) = \Phi(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}) - e^{\varepsilon}\Phi(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta})$, where $\Delta = \max_{\mathcal{D}\sim\mathcal{D}'} \|h(\mathcal{D}) - h(\mathcal{D}')\|_2$ is L2 sensitivity of h and Φ is CDF function of $\mathcal{N}(0, 1)$.

[2] Mathematically, the noisy aggregated output of Eq. 3 is expected to follow probability distribution (range from 0 to 1) as the mean of noises is 0. But some output values may be occasionally out of [0,1]. If so, we follow Tian et al. (2022) to re-normalize the out-of-bound value to 0 or 1. Practically, we observed (in Fig. 2) being out-of-bound is extremely rare since N is large ($3k$) and the first term dominates Eq. 3 as (Tian et al., 2022).

Method		3750 training samples				6000 training samples			
		P	R	F1	Acc	P	R	F1	Acc
Non-DP	Private	0.224	0.367	0.240	0.367	-	-	-	-
	DA w/ Public	0.280	0.347	0.280	0.348	0.313	0.347	0.287	0.348
DP($\varepsilon = 4$)	DP-SGD	0.130	0.310	0.170	0.314	-	-	-	-
	Ghost	0.143	0.289	0.166	0.291	-	-	-	-
DA w/ DP($\varepsilon = 4$)	DA w/ DP-SGD	0.290	0.350	0.277	0.352	0.283	0.350	0.280	0.352
	DA w/ Ghost	0.303	0.347	0.297	0.344	0.283	0.350	0.297	0.350
	Ours	**0.340**	**0.373**	**0.337**	**0.372**	**0.353**	**0.377**	**0.340**	**0.376**

Table 1: Main results comparing all the baselines on two size of datasets on text classification tasks.

Sensitivity Analysis of Knowledge Distillation (KD). We denote the output distribution of m-th teacher model is $P_m(\mathcal{Y}|x)$. The aggregation function $h(\mathcal{D}) = \sum_{m=1}^{M} P_m(\mathcal{Y}|x)$ is the summation of the output probability over all the teachers. Note that each teacher's data are disjoint (§3.4) and teachers have no duplicate samples, changing one sample only affects one teacher's output. For neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing only one sample, we denote i-th teacher is affected by the different sample in $\mathcal{D}$ and $\mathcal{D}'$. The output distributions $P_i(\mathcal{Y}|x)$ and $P_i'(\mathcal{Y}|x)$ (from $\mathcal{D}$ and $\mathcal{D}'$ respectively) are different. The sensitivity Δ_{KD} in Lemma 4 is (See deductions in App. A),

$$\begin{aligned}\Delta_{KD} &= \|h(\mathcal{D}) - h(\mathcal{D}')\|_2 \\ &\le \|P_i(\mathcal{Y}|x) - P_i'(\mathcal{Y}|x)\|_2 \le \sqrt{2}.\end{aligned} \tag{4}$$

Sensitivity Analysis of Tutor. In Eq. 3, each sample j's distribution $P_j(\mathcal{C}) = \{P_j(c_i)\}_{i=1}^{|\mathcal{C}|}$ is the binary-valued probability distribution, its value on the label is 1 while the value on all other categories is 0. $P_j(c_i) = \mathbb{I}(c_i \text{ is } x_j\text{'s label})$, where $\mathbb{I}$ is the indicator function. The function $g(\mathcal{D}) = \sum_j^N P_j(\mathcal{C})$ is the distribution on whole datasets. For neighboring dataset $\mathcal{D}$ and $\mathcal{D}'$, the different sample l in $\mathcal{D}$ and $\mathcal{D}'$ affects two terms: $P_l(\mathcal{C})$ in $g(\mathcal{D})$ and $P_l'(\mathcal{C})$ in $g(\mathcal{D}')$. The sensitivity Δ_{tutor} in Lemma 4 is (See deductions in App. B),

$$\begin{aligned}\Delta_{tutor} &= \|g(\mathcal{D}) - g(\mathcal{D}')\|_2 \\ &\le \|P_l(\mathcal{C}) - P_l'(\mathcal{C})\|_2 \le \sqrt{2}.\end{aligned} \tag{5}$$

Composition of KD and Tutor. Our whole DP learning algorithm is actually the combination of KD algorithm (§3.4) and tutor algorithm (§3.5). According to composition theorem for (ε, δ)-DP (Dwork et al., 2006), when the KD algorithm $\mathcal{M}_{KD}$ satisfies (ε_{KD}, δ_{KD})-DP and the tutor algorithm $\mathcal{M}_{tutor}$ satisfies (ε_{tutor}, δ_{tutor})-DP, then the combination (i.e. our whole algorithm) $\mathcal{M}$ satisfies ($\varepsilon_{KD} + \varepsilon_{tutor}$, $\delta_{KD} + \delta_{tutor}$)-DP. (See deduction in App. C). In summary, adding the noises according to in Lemma 4 to KD and the tutor is sufficient to preserve ($\varepsilon_{KD} + \varepsilon_{tutor}$, $\delta_{KD} + \delta_{tutor}$)-DP for the composition of KD and the tutor where the sensitivity is $\sqrt{2}$.

5 Experimental Settings

Datasets. We evaluate the methods on "Medical Transcriptions"[3] dataset, which is a dataset in the medical domain with medical transcription samples from 40 various medical specialties. It contains 5k items. Medical data are extremely hard to find due to HIPAA privacy regulations (Act, 1996). This dataset was scraped from mtsamples.com. We performed basic text processing on the data, converting all text to lowercase and removing punctuation. Subsequently, we randomly divided 75% of the samples for training and 25% for testing.

Evaluation Metrics. Metrics consists of: accuracy (Acc), precision (P), recall (R), and F1-score (F1, the harmonic mean of P and R).

Comparing Methods. We used two non-DP methods as the performance upper or lower bound: (1) Private directly trains on private data without protections. (2) DA w/ Public trains on data synthesized by public GPT-3.5 without accessing private data.

We use DP-based methods as: (1) DP-SGD (Abadi et al., 2016) trains on private data based

[3] www.kaggle.com/tboyle10/medicaltranscriptions

Method	3750 training samples			
	P	R	F1	Acc
Ours	0.340	0.373	0.337	0.372
Ours − Multi-teacher	0.280	0.297	0.213	0.298
Ours − Gaussian	0.327	0.333	0.263	0.335
Ours − Tutor	0.340	0.340	0.313	0.340
Ours − Tutor + Label Dist.	0.347	0.377	0.340	0.377

Table 2: Ablation studies. “+" or “−" means using or not using the given strategy.

on DP-SGD with noises on gradients. (2) `Ghost` (Li et al., 2022) trains on private data with DP-Adam and “ghost clipping”. (3) `DA w/ DP-SGD` trains DP-SGD on private data to select pseudo DA samples. (4) `DA w/ Ghost` trains Ghost on private data to select pseudo DA samples. (5) `Ours` denotes our proposed method. Note that `DA w/ DP-SGD` and `DA w/ Ghost` also imitates the noisy label distribution for a fair comparison with `Ours`. (See implementation details in App. D).

6 Experimental Results

6.1 Main Results

Table 1 presents the overall performance and we can observe that: without data augmentation (DA), `Private` acts as performance up-bound since it fully accesses the private data without any private protection. There are only 3750 train samples without DA. The Noisy-SGD methods (i.e. `DP-SGD` and `Ghost`) with DA exhibit significant increases to the same methods without DA, which shows the effectiveness of our designed DA framework. Note that we keep the sample number of DA and non-DA methods same for a fair comparison, and DA still works better since DA obtains high quality samples with a less privacy cost.

`Ours` outperforms other baseline methods in a meaningful range of privacy protection (ε is 4), even `Private`. The reason for outperforming `Private`, which acts as the up-bound, is that our synthesized data have higher quality and diversity than the private data since they are sampled from LLMs, which improves the generalization ability of classification models as training data are not sufficient. When the training samples selected from a fixed-size synthetic dataset increase from 3750 to 6000, the performance of all methods generally improves.

6.2 Ablation Studies

We conducted ablation studies to evaluate the effectiveness of our proposed components. Table 2 presents the precision, recall, F1-score, and accuracy of `Ours`’ variants on the downstream task. (1) `Ours − Multi-teacher`: only use one teacher to train the discriminator instead of multiple teachers, where its subpar performance is due to this teacher completely determines the prediction result when there is only one teacher model, and the noise will directly affect the prediction result and bring a great negative impact. where its subpar performance is due to when there is only one teacher model, this teacher completely determines the prediction result, and the noise will directly affect the prediction result and bring a great negative impact. (2) `Ours−Laplace`: use the Laplace mechanism to add noise to the discriminator instead of the Gaussian mechanism. The result demonstrates that the performance of using the Gaussian mechanism in our framework is better than using Laplacian noise. (3) `Ours − Tutor`: discard the tutor (with the label distribution) and simply select samples uniformly considering the label. The poor performance shows the significance of the proposed tutor and using label distribution in DA. (4) `Ours − Tutor + Label Dist.`: discard the tutor but directly imitate the private label distribution to verify the effectiveness of learning the private label distribution, which is confirmed by its excellent performance. Compared with `Ours`, it can learn a label distribution closer to real data without any privacy protection. Our tutor achieves similar performance even after adding noise.

6.3 Privacy-utility Tradeoff

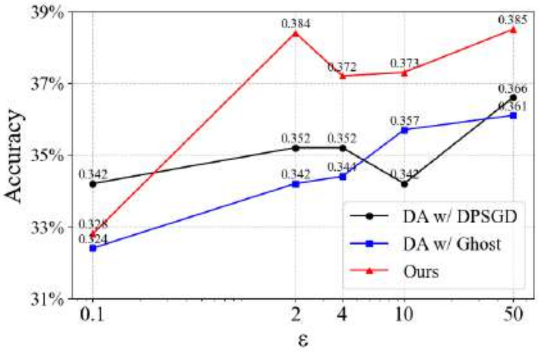

Figure 2: The private-utility tradeoff in accuracy across three DA w/ DP methods at varying ε. The vertical axis represents the accuracy on downstream text classification tasks.

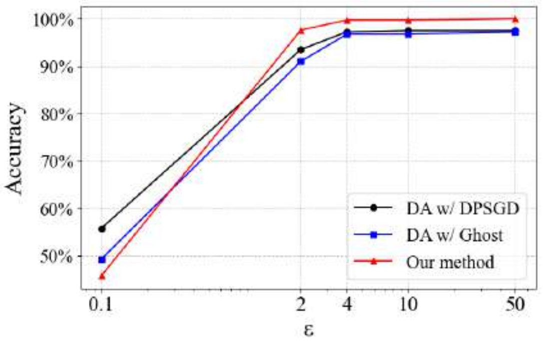

Figure 3: The private-utility tradeoff in the discriminator's accuracy across three DA w/ DP methods at varying ε. The vertical axis represents the accuracy on our constructed test set.

In Fig. 2 and Fig. 3, we show the private-utility trade-off curve of three DA w/ DP methods and their discriminators covering the range of meaningful protection (i.e. usually $\varepsilon \in [0.1, 10]$ (Triastcyn and Faltings, 2020)). In both figures `Ours` outperforms `DA w/ DP-SGD` and `DA w / Ghost` in most ranges of $\varepsilon \in [0.1, 10]$, showing that our method provides strong privacy protections while having excellent performance. As ε increases, privacy protection becomes weaker, leading to a gradual improvement in the performance of various methods. In Fig. 3, the Accuracy of the discriminator improves as ε increases, which aligns with our expectations. But when $\varepsilon > 2$, `Ours` exhibits fluctuations on downstream tasks. This is because the DP-based discriminator in `Ours` already achieves near-perfect accuracy (97.6%), and the performance of `Ours` has exceeded up-bound (36.7%). After exceeding the up-bound, the final performance is probably beyond the control of the discriminator (The discriminator judges whether the synthetic data meet privacy characteristics, but the performance of selected synthetic data by Ours exceeded the original privacy data).

6.4 Analysis On Teacher Numbers

To analyze the impact of teacher number on discriminator, we conducted experiments across varying teacher number with $\varepsilon = 4$ (in Fig. 4). `Teachers` denotes the aggregation of multiple teacher models; `Teachers + Noise` denotes to add noise to the voting results after aggregating multiple teacher models; `Student` denotes our final discriminator model. We conclude that: (1) As the teacher number increases, the performance of `Teachers + Noise` gradually improves. When the teacher number reaches 20, the impact of adding noise on the aggregated teachers' performance becomes negligible. (2) `Student`'s performance correlates positively with `Teachers + Noise`. (3) In most cases, `Student` outperforms `Teachers + Noise`. This is because `Teachers + Noise` reduces `Teachers`' original prediction accuracy by directly adding noise to the inference results. `Student` trains with noisy data, allowing the model to adapt to the noise.

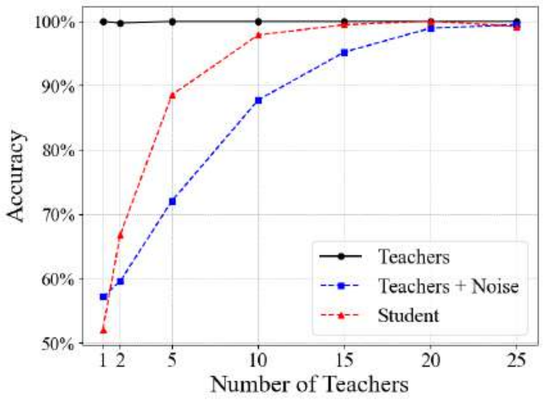

Figure 4: Analysis on teacher number. The vertical axis represents the prediction accuracy of the discriminator on our constructed test set.

6.5 Analysis On Tutor's Distribution

Fig. 5 (See App. E) illustrates two distinct label distributions: `Private Dist` represents the original private label distribution; `Tutor Dist` represents the label distribution after adding noise (discussed in §3.5), which provides stricter privacy protection by satisfying $\varepsilon = 0.4$. We observe a high consistency between `Tutor Dist` and `Private Dist`, which indicates that `Tutor Dist` effectively retains the characteristics of `Private Dist`.

7 Conclusion

In this paper, we proposed a DP-based DA method for text classification in private domains, which prompts a LLM to generate pseudo samples and uses a DP-based discriminator to examine the LLM's outputs. In this way, we transfer pseudo text generation task, which is a challenging DP paradigm, to a discrimination task. We construct a DP-based discriminator via knowledge distillation and construct a DP-based tutor to guide the sample generation with a low privacy cost. Theoretical analyses illustrates the bound of protections of our models and our experimental results shows our model's effectiveness.

8 Limitations

In our study, there were several limitations.

(1) First of all, we use GPT-3.5 instead of GPT-4 in our experiments and GPT-3.5 is not a latest and SOTA GPT API, which seems to limited the performance of our model. The main reason is that our baselines (Abadi et al., 2016; Li et al., 2022) are all based on GPT-3.5 for a fair comparison. In addition, compared to GPT-3.5, GPT-4 costs too much to obtain API keys. According to the official website, the fee for GPT 3.5 is 0.002$/1k Token, and the fee for GPT 4 is 0.06$/1k Token. In our actual baseline comparison, about 560,000 pieces of data were generated using GPT-3.5, and the total cost was about 300$. And if GPT-4 is used, the same number of tokens will cost more than 9000$. Therefore, the use of GPT-3.5 can not only effectively reduce the cost significantly, but also there will not be much difference in the generation effect. At present, we use GPT-3.5 for method verification, and GPT-4 will be used for effect verification in the future.

(2) Secondly, our approach is based on LLM-generated data, relatively dependent on the quality of the generated text. If the data are difficult to generate, or the overall quality of the generated text is poor, this may limit the advantages of our approach. Subsequently, we can leverage approaches such as prompt tuning, by carrying out a variety of excellent prompt engineering during the training phase of the LLM, enhancing the quality of text generated by LLM.

9 Ethical Considerations

We place significant importance on ethical considerations and adhere rigorously to the ACL Ethics Policy.

(1) This study introduces a novel text classification method, utilizing an LLM to generate pseudo samples and a DP-based discriminator to evaluate them, without ethical concerns regarding motivation or algorithmic approach, as no private information is used.

(2) Nevertheless, it's crucial to contemplate scenarios where individuals deliberately exploit our model for illicit purposes. For instance, someone might use the text generation model, used in this paper for generating pseudo data, to fabricate fake text or misinformation. This potential misuse poses a negative societal impact, using our model to generate false medical reports. Moving forward, we intend to implement constraints within our model to prevent the generation of texts for illegal activities, such as introducing filters to identify and flag potentially harmful or illegal content.

(3) Moreover, it's imperative to exercise caution in utilizing our model and refrain from assuming its infallibility. One potential unethical application involves gathering data from users who believe our model guarantees complete privacy protection, potentially overlooking the actual strength of privacy safeguards. This oversight could lead to adverse societal consequences. Therefore, we urge researchers intending to utilize this model to prioritize the efficacy of privacy protection. Additionally, measures should be taken to prevent researchers from collecting data from users who lack a proper understanding of our algorithm. For example, Clarify privacy policies and rules for data usage, and restrict access to collected data to authorized personnel only. We recommend that researchers ensure users contributing their data comprehend the risks associated with our model fully.

(4) In general, if the dataset contains privacy, it may be leaked during use. Regarding the datasets utilized in our experiments, the dataset generated by LLM does not contain the personally identifiable information of the real user. In contrast, the Medical Transcriptions dataset contains sample medical transcriptions for various medical specialties, allowing us to conduct experiments to assess the efficacy of privacy protection measures. It's important to note that the Medical Transcriptions dataset was previously made available to the public. Therefore, our research in this paper does not involve releasing any additional personal information of users.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *CCS*, pages 308–318.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*.

Borja Balle and Yu-Xiang Wang. 2018. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473.

Franziska Boenisch, Christopher Mühl, Roy Rinberg, Jannis Ihrig, and Adam Dziedzic. 2023. Individualized pate: Differentially private machine learning with individual privacy guarantees. *Proceedings on Privacy Enhancing Technologies*, 1:158–176.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. 2020. Deep learning with gaussian differential privacy. In *Harvard Data Science Review*, volume 2020. NIH Public Access.

Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. 2023. On the convergence and calibration of deep learning with differential privacy. *Transactions on Machine Learning Research*.

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

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2900–2904.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019a. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019b. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, pages 267–284.

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 *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised ner. *arXiv preprint arXiv:2010.01677*.

Jieren Deng, Chenghong Wang, Xianrui Meng, Yijue Wang, Ji Li, Sheng Lin, Shuo Han, Fei Miao, Sanguthevar Rajasekaran, and Caiwen Ding. 2022. A secure and efficient federated learning framework for nlp. *arXiv preprint arXiv:2201.11934*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Jinshuo Dong, Aaron Roth, and Weijie J Su. 2019. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *arXiv preprint arXiv:2305.15594*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014a. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Cynthia Dwork, Aaron Roth, et al. 2014b. The algorithmic foundations of differential privacy. In *TCS*, volume 9, pages 211–407.

Li Fei-Fei, Robert Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135.

Simson Garfinkel et al. 2015. *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. 2022. Exploring the limits of differentially private deep learning with group-wise clipping. *arXiv preprint arXiv:2212.01539*.

Zhiting Hu, Bowen Tan, Russ R Salakhutdinov, Tom M Mitchell, and Eric P Xing. 2019. Learning data manipulation for augmentation and weighting. *Advances in Neural Information Processing Systems*, 32.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Aaron Johnson and Vitaly Shmatikov. 2013. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Aeda: an easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Workshop in EMNLP*, pages 39–45.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. *Advances in neural information processing systems*, 30.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Xiang Li, Menglin Cui, Jingpeng Li, Ruibin Bai, Zheng Lu, and Uwe Aickelin. 2021. A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing*, 443:345–355.

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

Xiaotian Lin, Nankai Lin, Yingwen Fu, Ziyu Yang, and Shengyi Jiang. 2023. How to choose" good" samples for text data augmentation. *arXiv preprint arXiv:2302.00894*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Wakana Maeda, Yu Suzuki, and Satoshi Nakamura. 2016. Fast text anonymization using k-anonyminity. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, pages 340–344.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Nicolas Papernot, Martın Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11(12):255.

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9.

Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. 2009. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2021. Selective differential privacy for language modeling. In *arXiv preprint arXiv:2108.12944*.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, pages 245–248. IEEE.

Yu Suzuki, Koichiro Yoshino, and Satoshi Nakamura. 2018. A k-anonymized text generation method. In *Advances in Network-Based Information Systems: The 20th International Conference on Network-Based Information Systems (NBiS-2017)*, pages 1018–1026. Springer.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Zhiliang Tian, Yingxiu Zhao, Ziyue Huang, Yu-Xiang Wang, Nevin L Zhang, and He He. 2022. Seqpate: Differentially private text generation via knowledge distillation. *Advances in Neural Information Processing Systems*, 35:11117–11130.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Aleksei Triastcyn and Boi Faltings. 2020. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pages 9583–9592. PMLR.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. 2023. Privacy-preserving in-context learning for large language models. *arXiv e-prints*, pages arXiv–2305.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*.

Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. 2020. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *ICLR*.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2022. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. Seqmix: Augmenting active sequence labeling via sequence mixup. *arXiv preprint arXiv:2010.02322*.

Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. 2020. Private-knn: Practical differential privacy for computer vision. In *CVPR*, pages 11854–11862.

A Deduction of Sensitivity Δ_{KD}

We obtain the Equations (4) in §4 of the paper body since $P_i(\mathcal{Y}|x)$ and $P'_i(\mathcal{Y}|x)$ are the output distribution of i-th teacher model. where x is a given sample and $\mathcal{Y} = \{0, 1\}$.

$$\begin{aligned}\Delta_{KD} &= \|h(\mathcal{D}) - h(\mathcal{D}')\|_2 \\ &\leq \|P_i(\mathcal{Y}|x) - P'_i(\mathcal{Y}|x)\|_2 \\ &= \left(\sum_{j=1}^{|\mathcal{Y}|}(P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x))^2\right)^{1/2}\end{aligned} \tag{6}$$

We know $(P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x))^2$ is smaller than $|P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x)|$ since $|P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x)| \in (0, 1)$ for each j. Hence, we have,

$$\begin{aligned}&\left(\sum_{j=1}^{|\mathcal{Y}|}(P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x))^2\right)^{1/2} \\ \leq&\left(\sum_{j=1}^{|\mathcal{Y}|}|P_{ij}(\mathcal{Y}|x) - P'_{ij}(\mathcal{Y}|x)|\right)^{1/2} \\ \leq&\left(\sum_{j=1}^{|\mathcal{Y}|}|P_{ij}(\mathcal{Y}|x) + P'_{ij}(\mathcal{Y}|x)|\right)^{1/2}\end{aligned} \tag{7}$$

We know $|a + b| = a + b$ when $a, b \in (0, 1)$, so we have,

$$\begin{aligned}&\left(\sum_{j=1}^{|\mathcal{Y}|}|P_{ij}(\mathcal{Y}|x) + P'_{ij}(\mathcal{Y}|x)|\right)^{1/2} \\ =&\left(\sum_{j=1}^{|\mathcal{Y}|}P_{ij}(\mathcal{Y}|x) + \sum_{j=1}^{|\mathcal{Y}|}P'_{ij}(\mathcal{Y}|x)\right)^{1/2} \\ =&\left(1 + 1\right)^{1/2} \leq \sqrt{2},\end{aligned} \tag{8}$$

In summary, the upper bound of the sensitivity is,

$$\begin{aligned}\Delta_{KD} &= \|h(\mathcal{D}) - h(\mathcal{D}')\|_2 \\ &\leq \|P_i(\mathcal{Y}|x) - P'_i(\mathcal{Y}|x)\|_2 = \sqrt{2}.\end{aligned} \tag{9}$$

B Deduction of Sensitivity Δ_{tutor}

We obtain the Equations (5) in §4 of the paper body since $P_l(\mathcal{C})$ and $P'_l(\mathcal{C})$ are the sample l's distribution.

$$\begin{aligned}\Delta_{tutor} &= \|g(\mathcal{D}) - g(\mathcal{D}')\|_2 \\ &\leq \|P_l(\mathcal{C}) - P'_l(\mathcal{C})\|_2 \\ &= \left(\sum_{v=1}^{|\mathcal{C}|}(P_{lv}(\mathcal{C}) - P'_{lv}(\mathcal{C}))^2\right)^{1/2}\end{aligned} \tag{10}$$

Because l's distribution is a binary probability distribution with values 1 on its labels and 0 on all other classes, $\sum_{v=1}^{|\mathcal{C}|}(P_{lv}(\mathcal{C}) - P'_{lv}(\mathcal{C}))^2$ has a maximum value when sample l belongs to different categories in neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.

$$\begin{aligned}&\left(\sum_{v=1}^{|\mathcal{C}|}(P_{lv}(\mathcal{C}) - P'_{lv}(\mathcal{C}))^2\right)^{1/2} \\ \leq&\left(1 + 1\right)^{1/2} = \sqrt{2}\end{aligned} \tag{11}$$

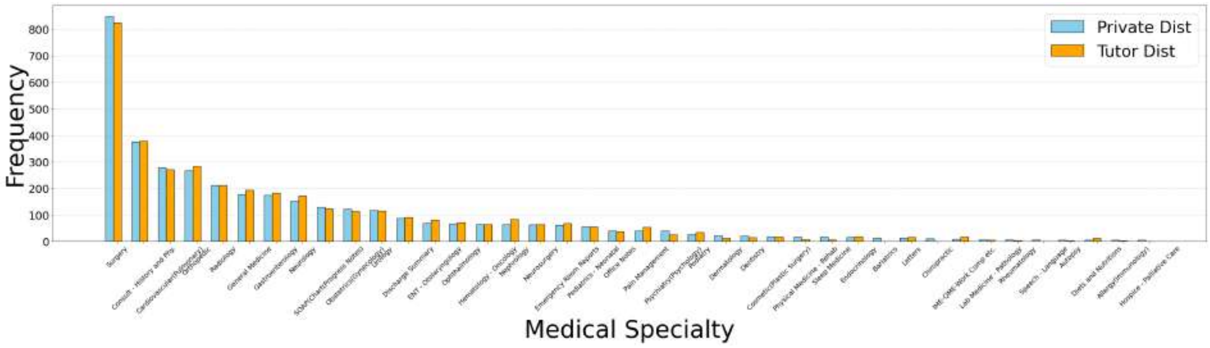

Figure 5: Label distributions. The horizontal axis enumerates all data labels, while the vertical axis represents the frequency of the labels.

In summary, the upper bound of the sensitivity is,

$$\begin{aligned}\Delta_{tutor} &= \|g(\mathcal{D}) - g(\mathcal{D}')\|_2 \\ &\leq \|P_l(\mathcal{C}) - P_l'(\mathcal{C})\|_2 \leq \sqrt{2}.\end{aligned} \tag{12}$$

C Detailed Deduction of Composition (KD and Tutor)

The KD algorithm $\mathcal{M}_{KD}$ satisfies (ε_{KD}, δ_{KD})-DP and the tutor algorithm $\mathcal{M}_{tutor}$ satisfies (ε_{tutor}, δ_{tutor})-DP. Since the output results of KD and Tutor are independent, for $\forall(\mathcal{D}_{KD} \times \mathcal{D}_{tutor}) \in (\mathcal{R}_{KD} \times \mathcal{R}_{tutor})$, there is

$$\begin{aligned}
&\Pr[(\mathcal{M}_{KD}), (\mathcal{M}_{tutor}) \in (\mathcal{D}_{KD} \times \mathcal{D}_{tutor})] \\
=&\Pr[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] \Pr[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}] \\
\leq&([e^{\varepsilon_{tutor}} Pr'[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}]] \wedge 1 + \delta_{tutor}) \\
&\times \Pr[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] \\
\leq&([e^{\varepsilon_{tutor}} Pr'[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}]] \wedge 1) \\
&\times \Pr[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] \\
&+ \delta_{tutor} \Pr[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] \\
\leq& e^{\varepsilon_{tutor}} Pr'[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}] \\
&\times Pr[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] + \delta_{tutor} \\
\leq& e^{\varepsilon_{KD+tutor}} Pr'[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}] \\
&\times [Pr'[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] + \delta_{KD}] + \delta_{tutor} \\
\leq& e^{\varepsilon_{KD+tutor}} Pr'[(\mathcal{M}_{tutor}) \in \mathcal{D}_{tutor}] \\
&\times Pr'[(\mathcal{M}_{KD}) \in \mathcal{D}_{KD}] + \delta_{KD} + \delta_{tutor} \\
=& e^{\varepsilon_{KD+tutor}} \\
&\times \Pr'[(\mathcal{M}_{KD}), (\mathcal{M}_{tutor}) \in (\mathcal{D}_{KD} \times \mathcal{D}_{tutor})] \\
&+ \delta_{KD} + \delta_{tutor}.
\end{aligned} \tag{13}$$

Therefore, it can be concluded from the above derivation that the combination $\mathcal{M}$ satisfies ($\varepsilon_{KD} + \varepsilon_{tutor}$, $\delta_{KD} + \delta_{tutor}$)-DP.

D Implementation Details

We select "gpt-3.5-turbo" with maximum 4,096 tokens construct synthetic text. In the DP-based DA. We fine-tune teachers and students based on the public pre-trained RoBERTa-large[4] (Devlin et al., 2019), which is representative and performs well in binary classification tasks. We set default number of teacher models is 15 and use the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 2×10^{-5}. The default target ε is 4 in §3.4 and 0.4 in §3.5, both δ is 10^{-6}. We use the autoDP[5] to obtain the standard deviation σ of the Gaussian mechanism (Dong et al., 2019) is 6. In the downstream classification task, we use ClinicalBERT(Huang et al., 2019) as the PLM with batch size of 64 and the AdamW optimizer with the learning rate of 10^{-6}. ClinicalBERT is a BERT derivative specifically for clinical medicine, outperforms RoBERTa-large in the medical field. All experiments were performed using a single NVIDIA RTX 3090 GPU.

E Label Distributions

Fig. 5 plots two distinct label distributions: `Private Dist` represents the original private label distribution; `Tutor Dist` represents the label distribution of tutor.

[4] We have tried LLM (i.e. GPT-3.5) as discriminator with the accuracy is less than 60%.

[5] https://github.com/yuxiangw/autodp