Differentially Private Language Models for Secure Data Sharing

Justus Mattern
RWTH Aachen
justus.mattern@rwth-aachen.de

Zhijing Jin
MPI & ETH Zürich
zjin@tue.mpg.de

Benjamin Weggenmann
SAP Security Research
benjamin.weggenmann@sap.com

Bernhard Schölkopf *
MPI for Intelligent Systems
bs@tue.mpg.de

Mrinmaya Sachan *
ETH Zürich
msachan@ethz.ch

Abstract

To protect the privacy of individuals whose data is being shared, it is of high importance to develop methods allowing researchers and companies to release textual data while providing formal privacy guarantees to its originators. In the field of NLP, substantial efforts have been directed at building mechanisms following the framework of local differential privacy, thereby anonymizing individual text samples before releasing them. In practice, these approaches are often dissatisfying in terms of the quality of their output language due to the strong noise required for local differential privacy. In this paper, we approach the problem at hand using global differential privacy, particularly by training a generative language model in a differentially private manner and consequently sampling data from it. Using natural language prompts and a new prompt-mismatch loss, we are able to create highly accurate and fluent textual datasets taking on specific desired attributes such as sentiment or topic and resembling statistical properties of the training data. We perform thorough experiments indicating that our synthetic datasets do not leak information from our original data and are of high language quality and highly suitable for training models for further analysis on real-world data. Notably, we also demonstrate that training classifiers on private synthetic data outperforms directly training classifiers on real data with DP-SGD.[1]

1 Introduction

Rapid advancements in the field of deep learning and natural language processing (NLP) have enabled companies, public institutions and researchers to extract information and gain knowledge from large-scale data generated by individuals. In many cases, it is desirable to share such data

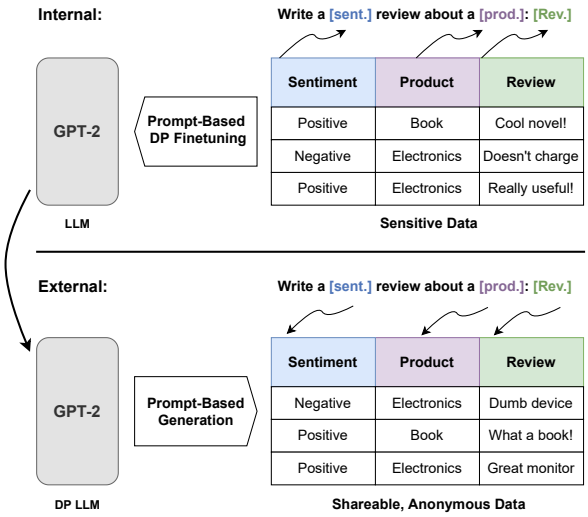

Figure 1: Main idea of our paper: To share potentially sensitive datasets with third parties, we train a language model (LM) on the sensitive data in a differentially private manner and consequently prompt the LM to generate synthetic samples with privacy guarantees.

with third parties, for example when analyses are performed by external consultants or in order to provide high quality benchmarks for the research community. This, however, entails a variety of risks related to privacy that cannot merely be solved by pseudonymization: A variety of deanonymization attacks enable the re-identification of individuals from tabular data such as movie ratings (Narayanan and Shmatikov, 2008), geolocation data (Lee et al., 2017) and notably also text (Koppel et al., 2009; Shrestha et al., 2017; Fabien et al., 2020). It is therefore highly desirable to develop anonymization mechanisms enabling secure data sharing, ideally with mathematical privacy guarantees as granted by differential privacy (DP) (Dwork and Roth, 2014).

Existing approaches anonymize every text sample individually by obtaining differentially private vector representations (Weggenmann and Kerschbaum, 2018; Fernandes et al., 2019) or using sequence-to-sequence approaches that rewrite a

* Equal Supervision.

[1] Our code is available at https://github.com/justusmattern/private-datasets-with-llms.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4860–4873
December 7-11, 2022

given sample to eliminate user-revealing information (Shetty et al., 2018; Feyisetan et al., 2019a, 2020a; Weggenmann et al., 2022), thereby following local differential privacy. As pointed out by Mattern et al. (2022), local DP requires a very high degree of noise which often leads to incoherent language and only little semantic overlap. The strict requirements of local DP are, however, not necessary if we assume that an entity aiming to share data already has access to the full collection of user-written texts and only wants to release an anonymized version of it.

In this paper, inspired by recent advances demonstrating the feasibility of training large language models (LLMs) in a differentially private manner (Li et al., 2021), we propose a globally differentially private data release mechanism relying on the generation of a "twin" dataset of the original, sensitive user data from large language models. As depicted in Figure 1, we train GPT-2 (Radford et al., 2019) to generate texts of our original dataset based on prompts inferred from the sample's individual attributes such as sentiment or topic. For fine-tuning, we use a differentially private optimization algorithm in order to protect the content of our training data. Subsequently, we sample from the trained model to generate a large number of synthetic, anonymous texts, resulting in a verifiably private "twin" dataset. We carefully evaluate our proposed method using popular NLP datasets such as IMDb movie reviews or Amazon product reviews. Here, we find that even after learning with strong privacy guarantees such as $\epsilon = 3$ or $\epsilon = 8$ from only a very limited amount of training samples such as 25 or 50, our generated data is of high quality and the classifiers trained on it achieve accuracies only $\sim$3% lower than those trained on the full original dataset containing thousands of samples. Notably, we also find that transformer based classification models trained on private data outperform models trained on real data with differentially private optimization. Finally, we show that the differentially private fine-tuning procedure effectively minimizes the risk of data leakage from language models that was previously discovered by Carlini et al. (2021).

2 Background

2.1 Differential Privacy

Differential privacy (DP) is a formal notion of privacy that is currently considered the state-of-the-art for quantifying and limiting information disclosure about individuals. It has been introduced by Dwork et al. (2006a) under the name *ϵ-indistinguishability* with the goal of giving semantic privacy by quantifying the risk of an individual that results from participation in data collection.

In the original, *central model* of DP, we consider *adjacent* datasets that differ by at most one record (i.e., one individual's data). A differentially private query on both databases should yield matching results with similar probabilities, i.e., answers that are probabilistically *indistinguishable*. This is achieved via random mechanisms that return noisy query results, thus masking the impact of each individual.

Definition 1. *Let $\epsilon > 0$ be a privacy parameter, and $0 \leq \delta \leq 1$. A randomized mechanism $\mathcal{M}$ on $\mathcal{X}$ fulfills (ϵ, δ)-*DP *if for any pair of adjacent inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, and all sets of possible outputs $Z \subset \operatorname{supp} \mathcal{M}$,*

$$\Pr\left[\mathcal{M}(\boldsymbol{x}) \in Z\right] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{M}(\boldsymbol{x}') \in Z\right] + \delta \,. \tag{1}$$

In the *local model* (Duchi et al., 2013), noise is added locally at the data source, before the data is collected and stored in a central database. A basic example is randomized response (Warner, 1965), where each survey participant either provides a truthful or a random answer depending on the flip of an (unbiased) coin. The local model makes the strong assumption that any two inputs are considered adjacent, which often makes it difficult to achieve a satisfying privacy-utility trade-off.

2.2 Differentially Private Optimization

An important application of DP is privacy-preserving machine learning to protect the privacy of the training data. Typically, neural networks are trained by optimizing a loss function using stochastic gradient descent (SGD) or a derived method such as Adam (Kingma and Ba, 2015), which iteratively compute gradients of the loss function over batches of samples from the training dataset. As shown by Song et al. (2013a); Bassily et al. (2014a); Abadi et al. (2016a), it is possible to implement a differentially private version of SGD (DP-SGD) by clipping the gradients and applying the Gaussian mechanism (Dwork and Roth, 2014): The latter works by applying noise from an isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where the standard deviation σ is derived based on the desired privacy parameters ϵ and δ.

To achieve good privacy-utility trade-offs, it is important to accurately track the total privacy budget spent throughout the entire training. In the context of DP, repeated executions of the same (here: Gaussian) mechanism is referred to as *composition*. Basic (Dwork et al., 2006b) and various more refined, advanced *composition theorems* (Dwork et al., 2010; Dwork and Rothblum, 2016; Bun and Steinke, 2016) have been stated in the literature that aim at providing tight bounds for the overall privacy budget. However, these advances still resulted in relatively loose bounds and thus large overall privacy budgets over the course of highly iterative algorithms such as DP-SGD. Tight worst-case bounds for composition were derived by Kairouz et al. (2015), however, it was shown to be computationally infeasible to compute them in general (Murtagh and Vadhan, 2016).

For this reason, specific efforts have been made to find tighter bounds and accurate approximations for the overall privacy loss: A first example that provides substantial reduced upper bounds is the moments accountant (Abadi et al., 2016a), which is closely related to Rényi DP (Mironov, 2017), a generalization of DP based on Rényi divergence. Gaussian and f-DP (Dong et al., 2019) provide an approximation of the total budget using the central limit theorem (CLT). Finally, Gopi et al. (2021); Koskela et al. (2020), inspired by Sommer et al. (2019), are able to compute the exact budget numerically up to arbitrary precision by aggregating the *privacy loss random variable* with fast Fourier transform.

3 Approach

We consider the following scenario to motivate our approach: an entity wants to implement NLP pipelines to gain insights from internal data, e.g., emails from customers. To seek advice and get support for modeling the data and building pipelines, the entity aims to share an excerpt of the internal data with a third party such as a consultant or a group of researchers. In order to do this without compromising the privacy of its customers, the aim is to synthesize a verifiably private "toy" dataset that reflects the properties of the original data without leaking private information. On such a toy dataset, a third party could research how to best solve the task at hand and train a model to perform inference on the actual internal data, without being able to access sensitive information about customers. Formally, we aim to achieve the following goal: We consider a dataset consisting of a training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$. Given $\mathcal{D}_{\text{train}}$ or a subset of it, we want to train a generative model to synthesize a dataset $\widetilde{\mathcal{D}}_{\text{train}}$ that does not leak information from the original $\mathcal{D}_{\text{train}}$. Furthermore, the synthesized dataset should share statistical properties with the original one so that a classification model trained on $\widetilde{\mathcal{D}}_{\text{train}}$ performs as well as if it was trained on $\mathcal{D}_{\text{train}}$ when making predictions about $\mathcal{D}_{\text{test}}$.

To achieve this, we use the pretrained autoregressive transformer model (Vaswani et al., 2017) GPT-2 (Radford et al., 2019) and use natural language prompts to enable the conditional generation of text based on desired textual attributes such as its sentiment, domain or genre provided in the prompt. Furthermore, we introduce a new training objective that penalizes the generation of samples fitting another label to reduce the risk of faulty labeled samples in our synthetic dataset. Finally, we fine-tune our model using a differentially private optimizer to provide privacy guarantees for our training data and to prevent information leakage from our model when subsequently sampling our synthetic dataset.

3.1 Conditional text generation with natural language prompts

As we want to control specific textual attributes of our synthetic data, we need to train our model in a manner that allows us to generate different types of texts corresponding to the desired attributes or labels present in our dataset. We consider a text sample to correspond to a set of M attributes of interest, namely $A := \{a_1, a_2, \ldots, a_M\}$, where each attribute a_j can take on a set of categorical values C_j. In the case of product reviews, a_1 could be the sentiment of a review that can take on the values $a_1 \in C_1 = \{\text{Positive}, \text{Negative}\}$ and a_2 can be the product category, so that $a_2 \in C_2 = \{\text{Books}, \text{Electronics}, \text{DVD}, \text{Kitchen}\}$. Our goal is to learn a model $p(x|a_1, ..., a_M)$ in order to controllably synthesize text samples according to our desired attributes.

A straightforward approach to realize this would be to train a single generative model for all possible attribute value combinations. This approach is, however, highly memory-intensive, as it requires us to store the weights of a large number of models that grows exponentially with the number of categorical attributes. Following recent work

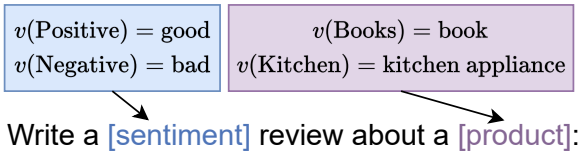

Figure 2: Our template-based approach for generating task instructions. A template consists of placeholders for verbalizations of different attribute values.

(Schick and Schütze, 2021a), we therefore train a single language model to conditionally generate texts based on task instructions. Beyond reducing our memory needs, this approach allows us to leverage our model's pretraining knowledge and to perform text generation with only very little training samples (Schick and Schütze, 2021a). Our instructions $\boldsymbol{i}(a_1, .., a_M)$ are formed using a template with placeholders that are filled out with verbalizations $v(a_j)$ taking on different forms for different values of every attribute a_j. An example of such an instruction template is visualized in Figure 2.

During the training stage, we use a differentially private optimizer to fine-tune our language model to generate each text sample within the original dataset based on the prompt corresponding to its individual attributes. Subsequently, we can synthesize a new dataset by controllably sampling text based on our desired attributes passed in the prompt. To generate a private "twin" dataset, one might use the same distribution of textual attributes as in the original dataset. Alternatively, the instruction-based approach allows us to control and change such ratios, for instance if we desire to debias our original data.

3.2 Reducing faulty labels with prompt-mismatch objective

The standard training objective for autoregressive language modeling is to minimize the negative log-likelihood (NLL) of every token given its previous tokens. We incorporate the natural language instructions (Radford et al., 2019; Brown et al., 2020) into this training objective. For every text sequence $\boldsymbol{x}$ and its corresponding attribute values $a := (a_1, ..., a_M)$, we construct the concatenated sequence $\boldsymbol{i}(a) \oplus \boldsymbol{x}$ which prepends a corresponding task instruction to each text sample. Let L denote the length of this concatenated sequence and let w_l be the sequence's l-th token. Our NLL loss is now

$$\mathrm{NLL}(\boldsymbol{i}(a) \oplus \boldsymbol{x}) = -\sum_{w_l \in \boldsymbol{i}(a) \oplus \boldsymbol{x}} \log p(w_l | w_{<l}) \; . \tag{2}$$

This objective encourages the model to generate correct samples for a given instruction. However, it does not minimize the likelihood of generating wrong samples corresponding to another prompt and therefore attribute. This is specifically unfavorable for our goal of generating synthetic training datasets as every generated text having an error of this kind corresponds to a wrongly labeled training sample. To address this, we extend the training objective with a term penalizing the generation of a given sample for a wrong prompt. Specifically, let I_{wrong} denote the set of all prompts not matching the given attribute values $a_1, ..., a_M$, specifically

$$I_{\text{wrong}} := \{\boldsymbol{i}(\overline{a}_1, ..., \overline{a}_M) \mid \overline{a}_j \in C_j \setminus \{a_j\}\} \; . \tag{3}$$

We now define the overall training loss we are aiming to minimize as

$$\begin{aligned} \mathcal{L}_{\text{ovr}} &= \mathrm{NLL}(\boldsymbol{i}(a) \oplus \boldsymbol{x}) \\ &- \frac{\lambda}{|I_{\text{wrong}}|} \sum_{i_{\text{w}} \in I_{\text{wrong}}} \mathrm{NLL}(\boldsymbol{i}_{\text{w}} \oplus x)) \; , \end{aligned} \tag{4}$$

where λ is the hyperparameter to balance the two losses. Note that in practice, when the number of possible labels is high, this computation might be inefficient and the objective too complex for the model to realize. In this case, one might randomly sample a few class labels for the wrong prompt in every training batch or penalize the generation for class labels that are the most similar to the correct one.

4 Evaluation

We conduct extensive evaluation measuring the utility and privacy of our generated data as well as the quality of its language. In this section, we describe the datasets we use as well as our evaluation metrics, implementation details and results.

4.1 Datasets

We use two publicly available datasets that are widely used for evaluating the performance of text classification models:

Table 1: Accuracy of classification models trained on synthetic data.

	IMDb			Amazon					
	Sentiment			Sentiment			Product Category		
# Train Samples	25	50	5000	25	50	3000	25	50	3000
BERT:									
$\epsilon = 3$	82.8%	88.3%	89.1%	85.2%	87.2%	88.5%	98.6%	98.7%	98.9%
$\epsilon = 8$	86.0%	87.6%	89.1%	87.4%	85.9%	89.2%	98.5%	98.9%	98.9%
$\epsilon = \infty$	86.5%	87.6%	89.2%	89.2%	88.5%	89.2%	98.7%	98.8%	99.0%
TF-IDF:									
$\epsilon = 3$	71.7%	78.3%	81.0%	69.5%	75.4%	79.1%	96.8%	97.0%	98.0%
$\epsilon = 8$	76.4%	79.2%	82.6%	74.9%	74.5%	78.3%	96.8%	98.2%	98.2%
$\epsilon = \infty$	80.2%	79.0%	82.5%	75.2%	77.9%	79.7%	97.6%	97.9%	98.1%

4.1.1 IMDb Movie Reviews

The IMDb movie review dataset[2] consists of movie reviews written by various authors. We use the two binary sentiment labels as attributes to condition our model on and use a random subset of 5,000 reviews for training and evaluation each.

4.1.2 Amazon Multi Domain Reviews

The Amazon multi domain review dataset was introduced by Blitzer et al. (2007) and consists of two thousand product reviews from each of the four product categories books, DVDs, electronics and kitchen appliances. Both binarized sentiment labels and the product categories books and electronics serve as attributes we consider. Our resulting training data consists of 3,000 training samples and 1,000 test samples.

4.2 Implementation Details

We implement and train our language models using the PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020) libraries and the 1.5B parameter implementation of GPT-2 (Radford et al., 2019). To fine-tune the language models, we employ the "privacy engine" of the `private-transformers`[3] package by Li et al. (2021). In line with their experiments, we also use DP-Adam (Dong et al., 2019; Bu et al., 2020), a differentially private version of the Adam (Kingma and Ba, 2015) optimizer. The privacy engine allows us to specify desired target privacy parameters ϵ and δ, from which the standard deviation parameter σ for the Gaussian mechanism is derived using either Rényi DP (Mironov, 2017), the CLT (Dong et al., 2019), or the FFT accountant (Gopi et al., 2021). Following Li et al. (2021), we set $\delta = \frac{1}{2*|D_{train}|}$ and vary the parameter ϵ in our experiments. To obtain reliable results for training our generative models on small subsets of the training samples, we sample three random subsets for every size and report averaged results from these three experimental runs. We trained GPT-2 over five epochs when using a differentially private optimizer and merely two epochs when using a non-private optimizer, as the latter tended to overfit quickly on the small training set. To further mitigate this, a smaller learning rate turned out to be more effective for non-private optimization: While we used a learning rate of 8e-6 with DP-Adam, we obtained the best results for non-private optimization with a learning rate of 5e-7. Lastly, we chose the hyperparameter $\lambda := 0.2$. We generated our synthetic datasets using the original distribution of sentiment and product category attributes, which was 50 / 50 in all cases. To sample from GPT-2, we use nucleus sampling (Holtzman et al., 2020) with $p = 0.8$ across all experiments. Our results were not obtained through an extensive hyperparameter search but educated guesses over a couple of iterations to avoid large computational effort. All experiments were performed using a NVIDIA Tesla A100 GPU. With this setup, a training epoch over 1,000 text samples took approximately five minutes.

Table 2: Accuracy of classification models trained on real data.

	IMDb	Amazon	
	Sentiment	Sentiment	Product
BERT:			
$\epsilon = 3$	83.6%	79.5%	95.2%
$\epsilon = 8$	86.7%	83.4%	96.6%
$\epsilon = \infty$	90.9%	91.2%	98.9%
TF-IDF:			
$\epsilon = \infty$	85.5%	75.8%	98.6%

[2] https://datasets.imdbws.com/

[3] https://github.com/lxuechen/private-transformers

4.3 Experimental Results

As stated previously, we aim to synthesize datasets that (1) reflect properties of the original data and can be used to train classifiers that perform similar to those trained on the original data, (2) are private and do not leak information from the original data and (3) are diverse and of high language quality. Accordingly, we perform experiments with metrics maesuring these attributes and report our results in the following:

4.3.1 Data Utility

To measure the utility of our datasets, we train classification models for each attribute on both our original data and the generated data and compare their performances when evaluating them on our real test data. Ideally, our anonymized twin datasets should lead to classifiers that are as accurate as those trained on our original data. To account for various settings including those with computational constraints, we train a shallow support vector machine classifier based on Tf-idf encodings as well as a deep BERT (Devlin et al., 2019) based classifier with 110M parameters. Furthermore, as an interesting baseline, we evaluate the performance of BERT trained on real data with differentially private optimization using the code from Li et al. (2021). The classification accuracies for models trained on synthetic data are shown in Table 1 and classification accuracies for models trained on real data are shown in 2. In the following, we summarize our key findings:

Synthetic data is almost on par with real data: The performance of classification models trained on generated data only drops minimally compared to those trained on real data. Across all three classification tasks, for both datasets with privacy budgets of $\epsilon = 3$ and $\epsilon = 8$, the accuracy of BERT and Tf-Idf based models is never less than 3% of the accuracy obtained for real data when given access to all training samples (5,000 and 3,000 for IMDb and Amazon, respectively).

Classifiers trained on synthetic data outperform private classifiers trained on real data: Notably, when comparing the results of transformer based classifiers trained on our synthetic data (Table 1) to those trained on real data with differentially private optimization (Table 2), we find that the former substantially outperforms the latter across all tasks for both $\epsilon = 3$ and $\epsilon = 8$. This raises the question whether the intermediate step of private data generation should always be performed rather than training classifiers with DP-SGD.

Private data generation shows high utility in few-shot settings: Lastly, even when given only as little as 25 or 50 samples, GPT-2 can generate datasets that lead to high-performing classifiers, which can most likely be attributed to the utilization of pretraining knowledge through our prompting techniques. Therefore, beyond the anonymization of existing datasets, our method can be used to enlarge existing small datasets in a private manner.

4.3.2 Data Privacy

To the best of our knowledge, methods aiming to measuring the privacy of textual data are an active area of research (Carlini et al., 2021; Brown et al., 2022) and there is no standardized and agreed upon way to do so. In our experiments, we follow Carlini et al. (2021) by counting the number of instances in which our synthetic dataset contains samples that are extremely close to a sample from the training data and can therefore be considered a duplicate: For every sample $\boldsymbol{x}_i$ from our training data used for the language model and every $\boldsymbol{x}_j \in \widetilde{D}_{\text{train}}$, we measure the set of trigrams $\text{g}_3(\boldsymbol{x}_i)$, $\text{g}_3(\boldsymbol{x}_j)$. We consider the two samples as duplicates if

$$|\text{g}_3(\boldsymbol{x}_i) \cup \text{g}_3(\boldsymbol{x}_j)| \geq 2 * \min(|\text{g}_3(\boldsymbol{x}_i)|, |\text{g}_3(\boldsymbol{x}_j)|)$$

As we hypothesize that duplicates are relatively rare, we double the generated data compared to our utility experiments and search for them within 10,000 and 6,000 samples generated for the IMDb and Amazon dataset, respectively. Our results are depicted in Table 3 and demonstrate the significant reduction of data leakage from privately trained models.

	IMDb			Amazon		
# Samples	25	50	5000	25	50	3000
$\epsilon = 3$	1	0	1	0	0	0
$\epsilon = 8$	0	6	2	4	1	0
$\epsilon = \infty$	8	23	13	16	30	9

Table 3: Number of duplicates from the training data generated by language models

4.3.3 Language Quality

As a metric measuring the quality of our generated samples, we use the Mauve[4] (Pillutla et al.,

[4] https://github.com/krishnap25/mauve

2021) score to compute the similarity of the distributions of $\mathcal{D}_{\text{train}}$ and the generated data $\widetilde{\mathcal{D}}_{\text{train}}$ from every trained model. As can be seen in Table 4, higher ϵ values tend to increase the quality measured by Mauve, but overall seem not to be highly significant. As a reference, the Mauve score computed when comparing $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are 0.95 for IMDb and 0.94 for Amazon. Based on manual inspection, the quality of generated texts seems to be very high. Mismatches between prompts and generated texts (e.g. a negative review generated for a positive prompt) as well as incoherent generations do occur, but very rarely. Excerpts of the generated data can be seen in Table 5, failure cases can be found in Table 6 and 7 in the appendix.

	IMDb			Amazon		
# Samples	25	50	5000	25	50	3000
$\epsilon = 3$	0.81	0.83	0.81	0.82	0.81	0.83
$\epsilon = 8$	0.82	0.81	0.81	0.82	0.81	0.82
$\epsilon = \infty$	0.81	0.85	0.84	0.84	0.85	0.82

Table 4: Mauve scores measuring the similarity of generated data and $\mathcal{D}_{train}$.

5 Related Work

5.1 Text Anonymization

Substantial efforts have been made to enable the privacy-preserving processing of textual data through both private textual vector representations and by transforming text into readable anonymous formats. Approaches from the former category either aim at obtaining term frequency vectors using differentially private mechanisms (Weggenmann and Kerschbaum, 2018; Fernandes et al., 2019) or by using deep learning methods with adversarial training objectives (Coavoux et al., 2018). In the work by Qu et al. (2021), various local DP mechanisms are explored to obtain private BERT representations.

Methods aiming at rewriting texts in a privacy-preserving manner range from rule-based approaches using human-engineered text perturbations (Mahmood et al., 2019; Bevendorff et al., 2019) as well as word replacements through the perturbation of individual word embeddings using differential privacy (Feyisetan et al., 2019b, 2020b) to deep learning based approaches leveraging sequence-to-sequence models. These sequence-to-sequence models can either incorporate adversarial objectives penalizing the generation of author-revealing information (Shetty et al., 2018; Xu et al., 2019) or integrate differential privacy in the text sampling process (Bo et al., 2021; Weggenmann et al., 2022; Mattern et al., 2022).

Notably, various papers proposing the integration of differentially private mechanisms in deep learning architectures (Krishna et al., 2021; Beigi et al., 2019a,b; Alnasser et al., 2021) have been shown to actually violate differential privacy (Habernal, 2021, 2022). While these works still represent important contributions due to their good empirical results, it should be noted that the design of NLP systems with DP guarantees is a task that is prone to errors and should be approached carefully.

5.2 Differentially Private Language Model Training

As generative language models have been shown to leak training data (Carlini et al., 2021) and the embeddings of discriminative models have been shown to contain sensitive information about a text's originator (Song and Raghunathan, 2020), differentially private optimizers such as DP-SGD (Song et al., 2013b; Bassily et al., 2014b) and DP-Adam (Abadi et al., 2016b; Kingma and Ba, 2014) have been applied to a variety of NLP tasks. Large-scale pretraining of BERT using DP-SGD has shown to reap comparable masked language modeling performance to non-private BERT (Anil et al., 2021). For the tasks of text classification and named entity recognition, good performance has been obtained with BERT and DP-SGD, but only with large privacy budgets of $\epsilon = 100$ or higher. Recently, it has been demonstrated that with the correct choice of hyperparameters and fine-tuning objectives aligned with the pretraining procedure, both generative and discriminative language models can achieve high performance in various tasks even with stricter privacy bounds (Li et al., 2021; Yu et al., 2022). An active area of research is concerned with the empirical evaluation of a language model's privacy (Brown et al., 2022) using methods such as membership inference attacks (Hayes et al., 2019).

5.3 Synthetic Data Generation:

Synthetic data generation with privacy guarantees using methods such as DP-GAN (Xie et al., 2018), PATE-GAN (Jordon et al., 2018) or various related approaches has successfully been applied for structured tabular or visual data (Torkzadehmahani et al., 2019; Neunhoeffer et al., 2020; Chen et al., 2020). Beyond these methods, DPRP (Differen-

Table 5: Exemplary generated samples from models trained with only 25 text samples. The texts were selected by picking a random sample from $\mathcal{D}_{test}$ and finding the most similar one according to Sentence-BERT (Reimers and Gurevych, 2019) within the generated data.

	Amazon (Category=Electronics, Sentiment=Positive)
Original	I'm not any expert in the computer world. But this product it is very easy to install. The range is very strong. I'm very happy with this product!!!
$\epsilon = 3$	I really like this. It works great. The speed is a bit slow. But, after I learned that this is how a good keyboard feels. I can't say enough good things about this keyboard.
$\epsilon = 8$	This is a good piece of equipment. It is small, has a high quality and a reasonable price. I used it to test the sound from my computer monitor. I am happy with the sound.
$\epsilon = \infty$	I got this game to play on my PSP as a little side project. I'm quite happy with it. The graphics are pretty good and the story line is pretty entertaining, as well.
	Amazon (Category=Books, Sentiment=Negative)
Original	Good Lord, people must be busier than I thought. If you haven't been allowed to even think about your future after retirement, then maybe you will get something out of this book. For the rest of us, this is drivel
$\epsilon = 3$	I have been looking for an awesome, fun, easy to read book, like a classic from when I was a kid. I am very disappointed. It is not a fun read. It really is a bore. I have read the sequel, and am also disappointed. This is not one of those things that you read to pass the time. Not a good choice for a first book.
$\epsilon = 8$	I would've done better with another book I read. The plot is interesting but the characters are bland and the setting is really only a backdrop to the action and plot. It's a shame as I enjoyed the book, but this isn't a good read.
$\epsilon = \infty$	This book was boring, boring, and boring. I have been thinking about getting a new copy of this book ever since I read it, but this one didn't work for me at all. Not a bad idea, just not my cup of tea.

tially Private Data Release via Random Projections) (Gondara and Wang, 2020) has been proposed as a model free alternative for releasing small private datasets that does not require training a generative model. For the domain of text, synthetic data generation techniques have predominantly been developed and evaluated without considering privacy guarantees (Anaby-Tavor et al., 2020; Schick and Schütze, 2021b). Merely the work presented by Bommasani et al. (2019) is similar to our paper, but does not provide any quantitative results about the experiments.

6 Conclusion

In this paper, we explored the generation of synthetic datasets from differentially private language models as a solution for publicly sharing textual data while protecting the privacy of users whose data is being shared. Our experiments show that synthetic data from differentially private language models is of high quality and is very well suited as training data for further tasks while significantly reducing the risk of leaking the original data. Our approach can be applied in a variety of use cases working with sensitive data. An interesting challenge for future work is the anonymization of multimodal datasets consisting of tabular, visual and text data.

Limitations

Privacy Guarantee While differential privacy provides a statistical privacy guarantee, one can not be certain that a differentially private language model does not leak any sensitive information. As seen in our experiments, the differentially private models did leak some of their training data, even if significantly less than the non-private ones. This can be a concern when dealing with training data containing names, telephone numbers or even passwords.

Synthetic Data Quality As shown in Table 6 and 7, our models did in rare cases produce incoherent language or text samples that did not fit the desired control attributes. This can limit the quality of the generated data.

Limits of Controllable Generation The controllability of multiple fine-grained textual attributes in text generation remains a difficult challenge Lyu et al. (2021). We therefore need to assume that our approach will become less accurate the higher the amount of textual attributes we want to consider.

Ethical Considerations

Data privacy is a highly important issue for the responsible deployment of machine learning solutions. With our work, we directly contribute to this field of research.

As our method relies on large pretrained language models, it should be noted that users deploying these technologies need to be aware of their undesirable, human-like biases (Sheng et al., 2019; Abid et al., 2021). Methods for reducing these harmful associations are actively being developed by the research community (Liang et al., 2021; Schick et al., 2021).

Acknowledgments

This material is based in part upon works supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) in the project *Trade-EVs II*, FKZ: 01MV20006A; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the John Templeton Foundation (grant #61156); by a Responsible AI grant by the Haslerstiftung; and an ETH Grant (ETH-19 21-1). Zhijing Jin is supported by PhD fellowships from the Future of Life Institute and Open Philanthropy, as well as the travel support from ELISE (GA no 951847) for the ELLIS program. We also thank OpenAI Researcher Access Program for granting our team credits to their API.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016a. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016b. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. 2021. Privacy preserving text representation learning using bert. In *Social, Cultural, and Behavioral Modeling*, pages 91–100, Cham. Springer International Publishing.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

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

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

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

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019a. I am not what i write: Privacy preserving text representation learning.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019b. Privacy preserving text representation learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 275–276, New York, NY, USA. Association for Computing Machinery.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108, Florence, Italy. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online. Association for Computational Linguistics.

Rishi Bommasani, Steven Wu, and Xanda Schofield. 2019. Towards private synthetic text generation. In *NeurIPS 2019 Machine Learning with Guarantees Workshop*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy?

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. 2020. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23).

Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. 2020. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. 2006a. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006b. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Cynthia Dwork and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*, volume 9. Now Publishers Inc., Hanover, MA, USA.

Cynthia Dwork and Guy N Rothblum. 2016. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020a. *Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations*, page 178–186. Association for Computing Machinery, New York, NY, USA.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020b. 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.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019a. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019b. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Lovedeep Gondara and Ke Wang. 2020. Differentially private small dataset release using random projections. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 639–648. PMLR.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. 2021. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11631–11642.

Ivan Habernal. 2021. When differential privacy meets nlp: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1528.

Ivan Habernal. 2022. How reparametrization trick broke differentially-private text representation learning.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, volume 2019, pages 133–152. De Gruyter.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1376–1385, Lille, France. PMLR.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Antti Koskela, Joonas Jälkö, and Antti Honkela. 2020. Computing tight differential privacy guarantees using FFT. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2560–2569. PMLR.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Wei-Han Lee, Changchang Liu, Shouling Ji, Prateek Mittal, and Ruby B. Lee. 2017. Blind de-anonymization attacks using social networks. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, WPES '17, page 1–4, New York, NY, USA. Association for Computing Machinery.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2116–2138, Online. Association for Computational Linguistics.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Jack Murtagh and Salil Vadhan. 2016. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125.

Marcel Neunhoeffer, Steven Wu, and Cynthia Dwork. 2020. Private post-gan boosting. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. *Natural Language Understanding with Privacy-Preserving BERT*, page 1488–1497. Association for Computing Machinery, New York, NY, USA.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650, Baltimore, MD. USENIX Association.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

David M Sommer, Sebastian Meiser, and Esfandiar Mohammadi. 2019. Privacy loss classes: The central limit theorem in differential privacy. *Proceedings on privacy enhancing technologies*, 2019(2):245–269.

Congzheng Song and Ananth Raghunathan. 2020. *Information Leakage in Embedding Models*, page 377–390. Association for Computing Machinery, New York, NY, USA.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013a. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. 2013b. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Benjamin Weggenmann and Florian Kerschbaum. 2018. SynTF: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 305–314, New York, NY, USA. Association for Computing Machinery.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

A Failure Cases

We demonstrate manually selected failure cases in Tables 6 and 7. Generated texts from Table 6 do not reflect the attributes passed in the prompt accurately and would therefore correspond to wrongly labeled samples in our dataset. Table 7 features generated texts that contain logical or language errors.

Table 6: Failure cases in which the generated text does not fit the desired attributes. Model mistakes are marked in red.

# Train Samples	ϵ	Desired Attributes	Text
25	3	Category=Books Sentiment = Negative	I have read and enjoyed many of your books, but I can't recommend this one highly enough. I am not a big fan of vampires. They don't interest me at all This book, however, did. It was a fun, entertaining read. I loved the characters. I really didn't care if I was going to be able to go back in time to help save the future, or if I would be in for the long haul. There was a lot of tension throughout the book. I was glad I was able to finish it. I'm glad that I was not as upset with it as many others were. I gave this book a solid 5 stars.
50	3	Category=Books Sentiment = Negative	I was looking for a book on how to be a good leader and found it by chance. It seems that no matter how bad you have been, how bad things have gotten, or how terrible a person you are, a leader can still turn it around.
50	3	Category=Books Sentiment = Positive	I am not a huge fan of the Star Wars films but I really enjoyed this one. It is the first time I've ever seen the Star Trek movies and I enjoyed them as well. The plot is a bit different then you'd expect from a Star Trek story but I enjoyed it all the same.
50	8	Category=Electronics Sentiment = Negative	I have a brand new LG Electronics Vivid 5, and so far I'm very happy. The screen has a nice clear picture and has great picture quality, the speakers are good, the microphone is not great, the battery is adequate, and it's nice and light. This is the first of the Vivid series I have bought, and I'm glad I got one.

Table 7: Failure cases in which the generated text is incoherent or does not make sense. Model mistakes are marked in red.

# Train Samples	ϵ	Desired Attributes	Text
50	∞	Category=Electronics Sentiment = Negative	The device itself, while very good, is bad. It seems like the software doesn't like to write out its own ID (it should) so the app doesn't work. I've seen other users say the same thing and have had no problems. I haven't bought another bluetooth device, but I'm pretty sure it has something to do with it.
25	8	Category=Books Sentiment = Positive	I really enjoyed this book. I didn't get the full explanation as I am only a beginner but for beginners with the basic knowledge of a math and physics the theory is very sound. 2.0 stars.
4,000	∞	Category=Books Sentiment = Negative	This book is a compilation of some of my personal thoughts and ideas. The author's goal was to make me believe in witchcraft, but he completely failed.
50	8	Category=Electronics Sentiment = Positive	This is the best little portable guitar i've bought since the little portable juke box on a bike. The high quality quality quality of this product is great . I do wish i had bought a sooner the old jukebox but once you get it together its really nice to have.