Selective Metric Differential Privacy for Language Models

Anuar Maratkhan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Computer Science and Systems

University of Washington

2023

Reading Committee:

Martine De Cock, Chair

Anderson Nascimento

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington

Abstract

Selective Metric Differential Privacy for Language Models

Anuar Maratkhan

Chair of the Supervisory Committee:
Professor Martine De Cock
School of Engineering and Technology

Recent advancements in pre-trained language models (LMs) have led to many breakthroughs in Natural Language Processing (NLP). When applied for downstream tasks, such as text classifiers or chatbots, LMs can leak information about the large text corpora they were trained on. In privacy-preserving machine learning, it is common to apply Differential Privacy (DP) mechanisms that mitigate such leakage. The traditional notion of DP, where each record in the data is treated as sensitive, does not translate well to NLP tasks since some token sequences – such as addresses and social security numbers – may be sensitive while others are not. We introduce the new notion of Selective Metric Differential Privacy (SMDP) and a concrete mechanism to realize SMDP. To this end, we draw upon the recently proposed notions of Selective DP, in which records are treated as sensitive or not, and Metric DP, in which the notion of adjacent inputs is relaxed through the use of a metric. Our experiments show that GPT models trained on data privatized with our SMDP approach have higher utility than with Metric DP while preserving the same level of privacy protection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my advisors, Dr. Martine De Cock and Dr. Anderson C.A. Nascimento, for their guidance and support during my Master's study and research, for giving me academic freedom to lead this work and for technical discussions and feedback. I would also like to thank fellow students in the research group for their insightful discussions and feedback. Last but not least, I would like to thank my family for their continuous support, patience, and motivation during this journey.

DEDICATION

to my family.

Chapter 1
INTRODUCTION

Recent advancements in Natural Language Processing (NLP) enabled computers to process text as never before, bridging the gap between human-level and machine-level text processing performance. Language modeling is a fundamental component in many downstream applications, including text classification [28, 72, 47, 35], question answering [13, 40], machine translation [4, 41, 62], and summarization [61, 39, 26], to name a few. Large transformer architecture [67] based pre-trained language models such as BERT [13] and GPT [56, 57, 7] are what made NLP systems so pervasive. However, these language models (LMs) often require vast amounts of data, tens or hundreds of Gigabytes of text, and the size of the datasets used in NLP is growing rapidly [6].

Recent works show that language models have a tendency to memorize training data [8, 9, 18, 30, 65, 34, 33, 19, 49]. Even though most of the data that common LMs were trained on is made public, some datasets may still contain sensitive information like social security numbers, demographic information, or medical conditions of patients. Releasing these models to the public either as weights or in a black-box manner is unsafe because, as Carlini et al. [8], Nasr et al. [49], and other studies have demonstrated, LMs can leak sensitive information, even without any access to the model weights. Such leakage is potentially harmful to society and even illegal when it violates laws regarding the protection of personal data, such as the GDPR,[1] the CCPA,[2] and the AI Bill of Rights.[3]

---

[1]European General Data Protection Regulation
`https://gdpr-info.eu/`

[2]California Consumer Privacy Act
`https://oag.ca.gov/privacy/ccpa`

[3]The AI Bill of Rights, unveiled by President Joe Biden in October 2022, may become a law in the future.`https://www.whitehouse.gov/ostp/ai-bill-of-rights/`

The substantial uptake in the use of chatbots built on top of Large LMs made concerns regarding privacy even more prominent. A bug in ChatGPT, for instance, caused leakage of user conversation histories to other users [12]. OpenAI also reported that it had leaked users' first and last names, email addresses, payment addresses, and part of credit card numbers [36]. Use of ChatGPT for assistance in coding unintentionally leaked confidential corporate information [42, 50]. The fact that users' conversations are retained and used further to train the model [46] has led to discussions and concerns all over the world, from a temporary ban on the use of ChatGPT in Italy [43] over scrutinization in the E.U. [5, 11], to a privacy investigation in Canada [55]. At the time of writing, OpenAI uses data provided by users of ChatGPT for model training unless users explicitly opt out [60].

The perceived tension between the desire, on the one hand, to benefit from the remarkable capabilities of LMs and, on the other hand, wanting to protect the privacy of users can be eased through Privacy-Enhancing Technologies (PETs). Among the most widely adopted methods that mitigate private information leakage from trained machine learning (ML) models are mechanisms to achieve Differential Privacy (DP) [16]. Informally, DP ensures that the inclusion or exclusion of any record in a dataset is obscured in the sense that any output obtained from computations over the dataset would have been equally likely to be reached whether the record was present in the dataset or not. To achieve DP, algorithms to train ML models add noise to, e.g., the input data, the objective function, the gradients, or the model parameters. Within the NLP literature, Hu et al. [29] identify two main approaches of adding randomness to ensure DP: *gradient perturbation* methods, which add noise to gradients as in DP-SGD [1, 52, 63], and *embedding vector perturbation* methods, which add noise directly to embedding vectors before model training [24, 23, 21, 22, 32]. The method that we propose in this paper is of the latter kind and a generalization of Metric DP [23].

DP traditionally provides privacy protection for the worst-case scenario where each dataset record is treated as sensitive and should not be leaked. As Brown et al. [6] noted, this classical notion of DP is unsuitable for text data because not all tokens, words, or sentences are sensitive, and treating them as sensitive results in a considerable utility decrease. Ac-

knowledging that not all token sequences are equally sensitive means that there is potential to relax the classical notion of DP for NLP systems.

Recently, Shi et al. [63] moved in this direction by proposing Selective Differential Privacy (SDP), a notion of DP where attributes of a record in a dataset are defined as sensitive based on a policy function explicitly provided by the users or developers of the system. They present an algorithm to provide SDP, called Selective-DPSGD, in which the weight update procedure from DP-SGD (with clipping gradients and adding noise to gradients) is only applied to batches of records containing sensitive attributes, while for batches without sensitive attributes, the traditional weight update rule from SGD (Stochastic Gradient Descent) is applied.

In this paper, we leverage the idea of a policy function from SDP, which is a gradient perturbation approach, to design what is, to the best of our knowledge, the first embedding vector perturbation approach that acknowledges that not all tokens in the text are equally sensitive. To this end, we generalize the existing definition of Metric DP, or MDP for short, to the notion of "Selective Metric Differential Privacy" (SMDP), which guarantees the privacy of selected tokens (as in SDP) bounded by a privacy ε and a distance metric (as in MDP). We provide a corresponding mechanism to realize SMDP and train GPT-2 models on data privatized with SMDP and with MDP for varying values of ε. Our findings are that SMDP improves the utility (perplexity) of the GPT-2 models while preserving the same level of privacy protection.

As such, our contribution in this work is two-fold. First, we propose SMDP, a novel notion of DP well-suited for text-based applications, along with a mechanism to realize it (Chapter 4). Secondly, we perform a set of experiments and empirically compare our approach to MDP [23] in terms of utility on language modeling tasks and privacy evaluated using a well-known privacy attack [8] (Chapter 5).

Chapter 2
BACKGROUND

This chapter provides details on the background knowledge needed to understand the problem definition and the proposed solution. Namely, we discuss a language modeling task in more detail. Further, we present some of the most prominent privacy attacks used to quantify the privacy of language models, and the definition of Differential Privacy.

2.1 Language Models

Language models have become a common component in the pipeline for various NLP tasks like question answering, sentiment classification, and summarization. Language models learn probability distributions of tokens from text corpora in a self-supervised manner. The fundamental task in language modeling is *next token prediction*, which can be described by a statistical model:

$$Pr(w_1, \ldots, w_n) = \prod_{i=1}^{n} Pr(w_i \mid w_1, \ldots, w_{i-1}) \tag{2.1}$$

where $w_1, \ldots, w_n$ and $w_1, \ldots, w_{i-1}$ are sequences of tokens and $Pr(w_i \mid w_1, \ldots, w_{i-1})$ is the probability of the occurrence of token w_i based on the context $w_1, \ldots, w_{i-1}$.

Neural networks are state-of-the-art models for learning probability distributions from text. A language model, i.e. a neural network with parameters θ, is trained on large unlabeled corpora of text to maximize the likelihood function:

$$L(\theta) = \sum_{i=1}^{n} \log Pr(w_i \mid w_1, \ldots, w_{i-1}) \tag{2.2}$$

in which the probability on the right hand side is the conditional probability of token w_i

given the context $w_1, \ldots, w_{i-1}$, as predicted by the neural network with parameters θ.

The quality of language models is evaluated based on the learned associations of tokens (intrinsic evaluation) and downstream tasks (extrinsic evaluation). Intrinsic evaluation of language models is mainly done using a perplexity score. Perplexity $PP(W)$ measures the likelihood of token sequences and is defined as:

$$PP(w_1, \ldots, w_n) = 2^{-\frac{1}{n} \log_2 Pr(w_1, \ldots, w_n)} \tag{2.3}$$

which can also be simply written as:

$$PP(w_1, \ldots, w_n) = \sqrt[n]{\frac{1}{Pr(w_1, \ldots, w_n)}} \tag{2.4}$$

Perplexity describes how reasonable text is given that it is natural language text. The lower the perplexity, the more natural the sequence is. For instance, the sentence "The dog is barking." should have lower perplexity compared to the sentence "The bird is barking." because birds do not bark, and it is implausible that such texts exist.

2.2 Privacy Attacks

Since the growth of interest in machine learning research, recent studies have shown that machine learning models may leak sensitive data used in training those models. Some of the more popular approaches to privacy attacks are membership inference, reconstruction, and model extraction attacks.

Membership Inference Attacks. One of the most popular types of attacks in the machine learning community are membership attacks which were first proposed by Shokri et al. [64]. Membership inference attacks can have either black-box access to the target model, i.e., no access to the training set or model parameters, or white-box access. From the outputs of the target model, membership attacks try to identify whether a record x was part of the training data. Membership inference attacks are accomplished by differences in model responses to

training and out-of-training input samples. The lower the perplexity on the input sample, the higher the probability that it was part of the training dataset. To illustrate a membership attack, assume that an adversary has only black-box access to a model. The adversary gives inputs to the model and observes the outputs. The inputs themselves can be obtained by text generation or by selecting them from some datasets to which the adversary has access, and which may be different from the dataset that the model was trained on. Next the adversary computes the perplexity score of the outputs produced by the model. For samples that were used in the training set, the model will likely have lower perplexity (higher probability of occurrence). For an out-of-sample input the model will output higher perplexity (lower probability of occurrence) because it hasn't seen it in the training set. From this we can conclude that a sample with lower perplexity score was more likely to be used in training the target model.

Reconstruction Attacks. Reconstruction attacks aim to reconstruct the training data samples of the target model. Such attacks can reconstruct the full training sample or some partial information about the training sample. For instance, Carlini et al. [9] performed an attack on the GPT-2 language model in two phases. First, they applied GPT-2 for text generation tasks given some prefixes. Then, the authors ranked the generated samples based on six different configurations. As a result, the most successful setting could extract 67% of the training data.

Model Extraction Attacks. Model extraction attacks are primarily used as a stepping stone for the above attacks. Model extraction attacks generally assume a black-box access to the model. For instance, the attacker can only query the API of the ML-as-a-Service. The adversary uses such black-box access to extract extract information from the target model, and uses the collected information to reconstruct the target model [59]. For example, consider an adversary who collects the target model outputs and trains another model in a teacher-student fashion, also known as knowledge distillation. As a result, the trained model emulates the performance of the target model.

Other methods for model extraction include extracting the target model hyperparameters

[68], and model architecture, such as the number of layers, activation types, or optimization algorithm used in training the model [51].

2.3 Differential Privacy

Differential Privacy DP is a mathematical definition of privacy that aims to protect individual records in the data. Initially, DP ensured privacy in database systems where the two databases differing in one record produce the same output [15]. Later, Dwork and Roth [17])proposed an approach for differential privacy in machine learning. Further, Abadi et al. [1] adapted it for deep learning systems that use SGD.

For machine learning algorithms, DP ensures that the presence of any record in a dataset is obscured so that any output of the machine learning model trained on the dataset would have been equally likely whether the record was present in the dataset or not. The privacy provided in DP is controlled by the parameter ε, which corresponds to the privacy budget, and the parameter δ corresponds to the probability of violation of privacy. The smaller ε and δ are, the stronger the privacy guarantees.

Definition 1. (ε, δ)***-Differential Privacy*** *[15]. For privacy loss parameters* $\varepsilon \geq 0$ *and* $\delta \geq 0$*, a training algorithm* M *satisfies* (ε, δ)*-DP if and only if for any pair of training datasets* D *and* D' *that differ in only one record, and any set of possible output* $O \subseteq Range(M)$*:*

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

Therefore, the above definition of differential privacy guarantees a worst-case scenario of privacy leakage by treating any data record in the dataset D as sensitive and cannot be leaked. Consequently, such a definition of a privacy protection algorithm ensures universal protection against all types of adversaries compared to ad-hoc privacy protection techniques proposed in the literature, such as data sanitization and deduplication. However, the traditional definition of DP is strict and does not directly apply to language models due to the fuzzy boundaries of individual records [6].

Local Differential Privacy Unlike the traditional definition of DP from the Definition 1, Local Differential Privacy (LDP) does not need to trust a central authority that collects data from the users. With LDP, the users contributing data to the data collector can perturb their data locally before the collector can access it.

Definition 2. ***ε-Local Differential Privacy*** *[14]. For a privacy loss parameter $\varepsilon \geq 0$ a training algorithm M satisfies ε-LDP if and only if for any pair of input values $v, v' \in D$, and any set of possible output $O \subseteq Range(M)$:*

$$Pr[M(v) \in O] \leq e^{\varepsilon} Pr[M(v') \in O]$$

Chapter 3

RELATED WORK

There is a substantial amount of literature on providing *input privacy* in NLP, focusing on scenarios that arise when one wants to train a model over the data from multiple data holders who do not want to disclose their data to anyone or when a user Alice wants to use the trained model of a provider Bob for inference, but Alice does not want to disclose her input (e.g. text or prompt) to Bob, and Bob does not want to disclose his model parameters to Alice. Commonly used PETs to handle input privacy needs are Federated Learning (FL) [44] and Secure Multiparty Computation (MPC), see e.g. [58, 20, 2, 25].

Different from the above, in this paper, we are concerned with providing *output privacy*, i.e., mitigating leakage of information about the training data from a trained LM. Differential Privacy (DP) is broadly adopted in the privacy-preserving machine learning literature as the standard approach to provide output privacy. Below, we first list related work on the use of techniques to ensure the standard definition of DP (see Chapter 2) in NLP tasks, and then describe recent work on alternative definitions of DP that are more tailored towards the specific characteristics of textual data, and as such closer to our work.

3.1 Traditional DP in NLP

Since the DP-SGD algorithm for training neural networks with DP guarantees and its variations [1, 52] were proposed, this technique became a standard for using DP in deep learning, including in the privacy enhancement of LMs. For example, Carlini et al. [8], who proposed privacy attacks against LMs using canaries, train one of the target models with DP-RMSProp and empirically validate that differentially private training eliminates the model's memorization. Similarly, Jagannatha et al. [34] propose membership inference attacks on

ClinicalBERT (BERT trained on clinical notes) [31] and show that DP-SGD can reduce privacy leakage of their target models. Recent studies [73, 38] apply DP-SGD to fine-tune large pre-trained LMs and achieve high performance in terms of privacy-utility trade-off.

Some works propose adapting DP for LMs on a tokenizer level. For instance, Hoory et al. [27] present a DP WordPiece [70] algorithm with $\varepsilon = 1.1$. The DP WordPiece algorithm modifies the original WordPiece tokenization algorithm for BERT by adding Laplacian noise to the histogram of word counts. Their experiments using text extraction attacks from [8] in BERT trained on clinical notes indicate less memorization. Ponomareva et al. [53] propose a modification of Hoory et al.'s [27] approach for private SentencePiece [37] tokenization, providing sentence-level DP guarantees. Both works, in addition, apply DP-SGD to train BERT-like models after private tokenization.

Our work differs from the traditional DP in NLP literature because we propose a novel notion of DP that is more suitable for textual data.

3.2 Alternative Definitions of DP in NLP

Recently, Shi et al. [63] proposed a Selective DP method for LMs. Acknowledging that not all words in all contexts must be protected, Shi et al. [63] propose to apply DP-SGD only to batches with sensitive attributes. The latter is singled out by a policy function, which can be a simple heuristic that specifies that all numbers are sensitive or a more complex one based on a neural network trained to extract sensitive entities. Like all the other methods for DP in NLP that we have discussed so far, Selective DP [63] is a gradient perturbation method, unlike the embedding vector perturbation method that we propose in Chapter 4.

To relax the strict requirements of the canonical definition of DP, Chatzikokolakis et al. [10] propose Metric DP with distance metric d. Feyisetan et al. [24, 23] and Fernandes et al. [21, 22] propose to adopt Metric DP (MDP) to textual data with based on the Euclidean or Hyperbolic distance between embedding vectors. Metric DP applied to text is similar to local DP in the sense that each word w (or, technically, its embedding) is perturbed to provide plausible deniability. Such perturbation happens before model training. Local DP requires

that word w has a non-negligible probability of being transformed into any other word w', no matter how unrelated w and w' are, making it virtually impossible to enforce that the semantics of w is approximately captured by the privatized word w' [23]. MDP, on the other hand, gives a higher probability to words that are close to w and negligible probability to words in a completely different part of the vocabulary. The notion of "closeness", as captured by the distance function d, replaces the idea of neighboring datasets from traditional DP. The practice of perturbing embeddings to provide MDP has been further developed by, among others, Qu et al. [54] and by Xu et al. [71]. The latter uses the Mahalanobis (elliptical) norm to consider the shape of a particular space and improve the performance of private embeddings. Tang et al. [66] extend this framework by proposing a three-layer privacy protection mechanism. Their mechanism applies privatization from MDP [23] differently based on the type of words with different privacy levels.

Recently, Yue et al. [74] proposed Utility-optimized Metric Local DP (UMLDP) with privatization mechanisms SANTEXT and SANTEXT$^+$ that develop on the idea of MDP and Utility-optimized Local DP (ULDP) [48]. The SANTEXT$^+$ mechanism first divides the text into sensitive and non-sensitive token sets. Then, if the token is in the sensitive set, they sample substitutions based on MDP. If the token is non-sensitive, with probability $(1 - p)$, the token remains unchanged; otherwise, the mechanism samples new tokens with probability p.

Instead of distinguishing between sensitive and non-sensitive data, Wu et al. [69] propose an Adaptive Differential Privacy (ADP) without discrimination. In ADP, the authors estimate the probability of how private the token is based on the language model without any prior privacy information. Moreover, their proposed modification of Adam optimization adjusts the degree of DP noise injected into the model based on the privacy probability of a token. However, no formal definitions and proof that such a mechanism is differentially private were presented.

The method that we propose in Chapter 4 brings the idea of selective DP, which was originally proposed for gradient perturbation approaches [63] to the realm of embedding

perturbation approaches [23].

Chapter 4

METHODOLOGY

We first recall the formal notion of Metric DP (MDP) [10, 23], and then propose its generalization to Selective Metric DP (SMDP).

Definition 3. ***Metric Differential Privacy (MDP).*** *Given a privacy parameter $\varepsilon > 0$, a randomized algorithm $M : W \to W$ is called ε-Metric DP with distance function d if for any $w, w' \in W$ and $y \in W$:*

$$Pr[M(w) = y] \leq e^{\varepsilon d(w,w')} Pr[M(w') = y]$$

In Feyistan et al. [23] and subsequent works on MDP applied to text (see Hu et al. [29] and references therein), $M : X \to X$ where $X = W^l$ is the space of strings of length l with words in a dictionary W, i.e. the mechanism M converts a string into another string (the privatized string) by perturbing each word in the string. The insight that we build on below is that not all words are sensitive, and by being selective in the words we choose to perturb, we can create a privatized dataset of higher utility that still offers the same level of protection for the words considered sensitive. To this end, as in [63], we assume the existence of a user-defined **policy function** that denotes what is considered sensitive. Below, we assume such a policy function $F : W \to \{0, 1\}$ where $F(w) = 1$ means that w is sensitive, and $F(w) = 0$ that it is not.

Definition 4. ***Selective Metric Differential Privacy (SMDP).*** *Given a policy function F and a privacy parameter $\varepsilon > 0$, a randomized algorithm $M : W \to W$ is called (F, ε)-Selective Metric DP with distance function d if for any $w, w' \in W$ s.t. $F(w) = 1$,*

$F(w') = 1$, *and* $y \in W$*:*

$$Pr[M(w) = y] \leq e^{\varepsilon d(w,w')} Pr[M(w') = y]$$

The above definition provides plausible deniability for sensitive words only, e.g., $F(w) = 1$, $F(w') = 1$. The privacy guarantees for these words are the same as with MDP. Our novel notion, SMDP, does not guarantee privacy for the non-sensitive words.

We now propose a mechanism to implement SMDP in Algorithm 1 by adding lines 1, 5, and 6 to the MDP mechanism originally proposed by Feyisetan et al. [23]. The mechanism takes a word w, a policy function F, and a parameter $\varepsilon > 0$ as an input and outputs a privatized word y that will be further used for training a machine learning model. The mechanism further relies on a word embedding model $\phi : W \to \mathbb{R}^h$, where h is the dimension of the embedding vector. We use the Euclidean distance between two embeddings as our metric for SMDP. As for the embedding model, we chose GloVe because of its nature, which is independent of the words' context, which is, in contrast, different from transformer-based models, in which the words' embeddings are dependent on their context. Since our mechanism relies on finding the nearest words in embedding space, using context-dependent embedding models in our mechanism is non-trivial compared to context-free models like GloVe. Depending on the policy function F, our mechanism perturbs embeddings of the sensitive words by adding random noise N sampled from a distribution with density $p_N(z) \propto exp(-\varepsilon\|z\|)$ to the embedding $\phi(w)$ to obtain a noisy embedding vector $\hat{\phi} = \phi(w) + N$. As in [23], in the embedding space, we then look up the nearest word vector to the noisy embedding and replace the input word with this word. The embeddings of non-sensitive words remain unchanged, and therefore, the input word remains unchanged. Table A.1 in Appendix A presents samples from privatization mechanisms at different privacy levels.

The dictionary W in our approach depends on the choice of the embedding model since we rely on the model's vocabulary. To explain it further, we have to distinguish between two vocabularies. The first is of the embedding model (the dictionary of all words W in our case),

Algorithm 1 Selective Metric DP (SMDP) Mechanism

Input: Word w
Parameter: Policy function F, privacy parameter $\varepsilon > 0$, embedding ϕ
Output: Privatized word y

1: **if** $F(w) = 1$ **then**
2: Compute embedding $\phi(w)$
3: Perturb embedding $\phi(w)$ to get $\hat{\phi} \leftarrow \phi(w) + N$ with noise density $p_N(z) \propto exp(-\varepsilon\|z\|)$
4: Obtain perturbed word $y = argmin_{u \in W}\|\phi(u) - \hat{\phi}\|$
5: **else**
6: $y = w$
7: **return** y

and the second comes from the training dataset to be privatized. Considering that we do not want our privatization mechanism to replace unknown words with other random unknown words, we chose the embedding model's vocabulary as W. We then apply the mechanism to the known words only. As a result, the words present in the training set but not in the embedding model's vocabulary will remain unchanged so as not to hurt the model's utility.

Theorem 1. *For any policy function F and privacy parameter ε, the mechanism in Algorithm 1 satisfies (F, ε)-SMDP.*

Proof. To prove that the mechanism $M : W \to W$ satisfies (F, ε)-Selective Metric DP, we need to show that for two words $w, w' \in W$, the probability distributions over mechanism outputs $M(w)$ and $M(w')$ satisfy the following:

$$\frac{Pr[M(w) = y]}{Pr[M(w') = y]} \leq e^{\varepsilon d(w,w')} \tag{4.1}$$

The mechanism $M : W \to W$ perturbs sensitive words in lines 2-4 in the Algorithm 1, which corresponds to the mechanism from MDP [23]. Therefore, proof that the mechanism satisfies (F, ε)-Selective Metric DP for sensitive words follows the proof from [23]. As for non-sensitive words, the mechanism leaves them unchanged and, therefore, does not incur privacy loss for non-sensitive words. Hence, the mechanism satisfies (F, ε)-SMDP for both sensitive and non-sensitive words.

Chapter 5

EXPERIMENTAL RESULTS

5.1 Experimental setup

We hypothesize that SMDP improves the privacy-utility trade-off compared to MDP by improving the utility of the models on downstream tasks with the same level of privacy. To prove our hypothesis, we perform the following set of experiments. We chose a language modeling task on Wikipedia texts from WikiText-2 and WikiText-103 datasets [45] following the previous work [63] as it is a public dataset. The data splits in WikiText-2 are 600 articles for training, 60 for validation, and 60 for testing. WikiText-103 version of the dataset is comparatively larger (103 million tokens) and has 28,475 articles for training and 60 for both validation and testing. Similar to the previous work [63], we insert a canary "My ID is 145572." 10 times into the training set to evaluate privacy with a canary insertion attack [8], and treat all digits as sensitive in our policy function F. We evaluate the model's performance on the testing split of WikiText datasets. Note that both WikiText datasets share the same testing set.

We fine-tune a pre-trained GPT-2 small model with 124M parameters on the training set in three settings:

- **No DP**: The model is trained on the original training set without differential privacy.

- **MDP**: The models are trained on versions of the privatized training set using the mechanism from [23]. We create several privatized versions of the training set, corresponding to increasing privacy parameter ε values, using the same values for ε as in [23].

- **SMDP**: The models are trained on versions of the privatized training set using our approach in Alg. 1, using the same ε values as above.

Both No DP and MDP settings serve as a baseline for our approach.

5.2 Evaluation metrics

We evaluate all models using two metrics: a) perplexity for utility and b) canary exposure for privacy. **Perplexity** measures the likelihood of token sequences. Intuitively, the perplexity measures the surprise by the model on unseen data, and thus, the lower it is, the better. For a token sequence $X = (x_1, x_2, \ldots, x_t)$ the perplexity is defined as follows:

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i|x_{<i})\right\}$$

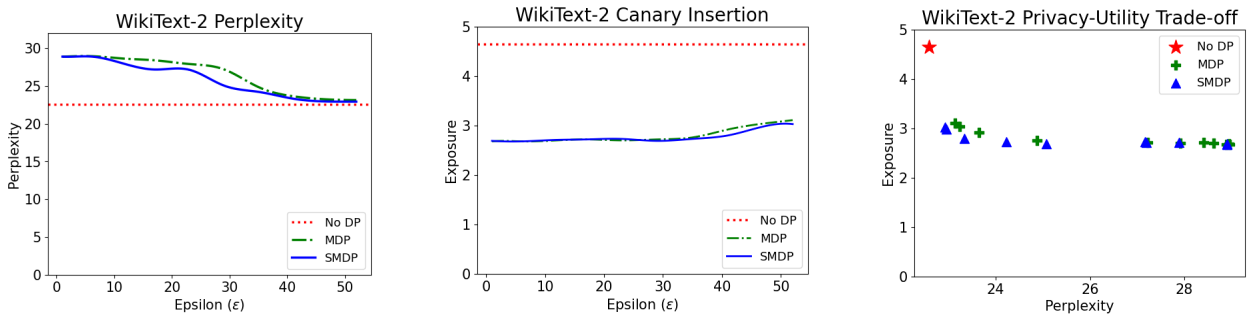
where $\log p_\theta(x_i|x_{<i})$ corresponds to the log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$ given model parameters θ.

The **exposure** measures how well the trained model guesses the inserted canary sequence [8]. Given a canary $s[r]$, the model parameters θ, and the randomness space $\mathcal{R}$, the exposure is based on the negative log-rank of $s[r]$ with constant $\log_2 \mathcal{R}$ and is defined as follows:

$$\mathbf{exposure}_\theta = \log_2 |\mathcal{R}| - \log_2 \mathbf{rank}_\theta(s[r])$$

In all our experiments, $\mathcal{R}$ is a randomness space of digits 0 to 9, and $|\mathcal{R}|$ is a constant. Considering that our inserted canary is of a format "My ID is xxxxxx", e.g., "My ID is 145572.", the randomness space $|\mathcal{R}|$ is $10^6 = 1,000,000$. Additionally, to enable fair evaluation of our approach, we calculate the negative log-rank of the canary based on the log-likelihood of a sequence "145572" given context "My ID is": $\log p_\theta(\text{"145572"}|\text{"My ID is"})$ since our policy function in all of our experiments protects digits only.

As explained above, we privatize the training dataset using the mechanisms from MDP and SMDP and then train models on the privatized data. The hyperparameters of the



(a) Perplexity (b) Canary insertion (c) Privacy-Utility Trade-off

Figure 5.1: Test set perplexity, canary insertion attack, and privacy-utility trade-off on WikiText-2.

models (learning rate, weight decay, warmup steps) were manually tuned to achieve the best perplexity scores on the validation set. The utility is further evaluated on the plain test set. We chose not to privatize validation and test sets to demonstrate a practical scenario in which the models are trained on sensitive data and evaluated on public data.

We used a pre-trained GloVe model trained on Common Crawl data with 840B tokens with a 2.2M vocabulary size for the privatization mechanism. We chose to utilize this model because it has a large vocabulary and is case-sensitive. Since GPT-2 models are all case-sensitive, we chose an embedding model with a case-sensitive vocabulary. The dimension of the embedding model used in our experiments is 300, e.g., the embedding model ϕ outputs 300-dimensional vectors.

5.3 Results

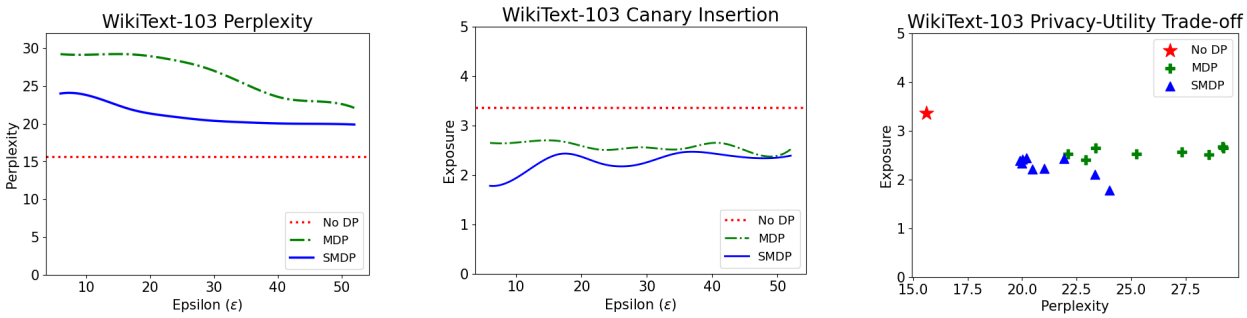
WikiText-2 Recall that a specific choice of ε depends on the metric space used in the mechanism and is not transferable across metrics. Additionally, the privacy parameter ε in MDP and SMDP is different from ε in traditional DP because, in MDP/SMDP, indistinguishability is bounded by ε times the distance between words. Therefore, in our experiments, the specific values of ε are chosen from the previous work [23]. We refer the readers to [23] for

further details on the choice of the privacy parameter ε.

Figure 5.1a shows models' utility measured by perplexity scores on the test set across different privacy levels. Recall that lower perplexity means higher utility and lower ε means higher privacy guarantees. The model's utility for No DP is represented by a red dotted horizontal line with the best perplexity score of 22.59, which serves as a baseline for the privatization mechanisms. The dash-dot green line represents the MDP baseline from [23]. The solid blue line is the model's utility for our approach, SMDP. Comparing MDP and SMDP approaches, the utility across all ε values is better for SMDP. For $\varepsilon = 29$, MDP achieves a 27.22 perplexity score, while our approach (SMDP) improves this score by 2.15, reaching 25.07 perplexity on the test set (see Table A.2 for detailed utility scores). Hence, with SMDP, we can achieve lower perplexity with the same privacy guarantees.

Figures 5.1b and 5.1c show the results for privacy evaluation experiments using canary insertion attack [8]. In Figure 5.1b, we can see the relation between various ε values (x-axis) and the success rate of the attack measured by the exposure metric (y-axis). The exposure values for MDP and SMDP overlap and achieve lower scores compared to No DP, which demonstrates the safety of the privatization mechanisms. Figure 5.1c shows the privacy-utility trade-off across three settings. The x-axis represents the models' perplexity on the test set, and the y-axis corresponds to the attack's success rate measured by exposure. Although No DP has lower perplexity (represented by a red star on the upper left of the plot), the exposure is worse compared to our approach (SMDP). If we compare MDP and SMDP performances, SMDP preserves the same level of privacy protection and achieves better utility. These results demonstrate that SMDP achieves better utility at the same level of privacy.

WikiText-103 Figure 5.2 shows results on WikiText-103. The utility for the SMDP approach is substantially better compared to MDP, improving perplexity by **5.15** on average. These results clearly demonstrate that SMDP improves utility significantly compared to MDP, leading to a smaller gap with No DP performance. We assume that privatization



(a) Perplexity (b) Canary insertion (c) Privacy-Utility Trade-off

Figure 5.2: Test set perplexity, canary insertion attack, and privacy-utility trade-off on WikiText-103.

mechanism performance is susceptible to such changes due to the growing size of the dataset. In our experiments, WikiText-2 has 2 million tokens in the training set, while the WikiText-103 version contains 103 million tokens. Hence, increasing the size of the dataset results in a higher number of perturbed words and shifts the original distribution of weights in the model further. Since SMDP perturbs digits only in our experiments, more words are left unchanged, which leads to more meaningful texts.

The results for privacy attack and privacy-utility trade-off are presented in Figure 5.2b and Figure 5.2c, respectively. We see a similar pattern for the canary insertion attack, with MDP and SMDP having less exposure than No DP. Additionally, SMDP exposure on WikiText-103 is better compared to both No DP and MDP across all epsilons, which shows that SMDP can be more private than the baseline [23]. The privacy-utility trade-off plot summarizes the results of utility and privacy experiments, demonstrating the superiority of our approach. The readers can find more detailed results in Appendix A.

We also observe that MDP requires more accurate hyperparameter tuning to prevent overfitting, e.g., higher weight decay (0.1) and lower learning rate (as low as 1e-9), especially on lower epsilons. If trained in a similar setting as SMDP, the models overfit significantly. In contrast, SMDP did not require such tuning, and we used hyperparameters close to the

No DP setting. This observation strengthens our assumption that MDP shifts the weight distribution substantially, while SMDP is more agnostic to such changes. Therefore, SMDP leads to much better results.

Chapter 6

CONCLUSION & FUTURE WORK

In this work, we propose a novel notion of Selective Metric Differential Privacy and a mechanism to realize it. Our approach relies on the notions of Selective DP [63], in which records are defined as sensitive or not based on a policy function, and Metric DP, in which the adjacency of inputs is relaxed through the use of a metric. Our experiments with WikiText-2 and WikiText-103 show that GPT models trained on the privatized data with the proposed SMDP approach achieve significantly higher utility compared to MDP with the same level of privacy protection.

Future work on SMDP may include a broader set of experiments, including more downstream tasks and privacy evaluations, various word embedding models, and distance metrics for privatization mechanisms. Although ε values in traditional DP and MDP/SMDP are incomparable, evaluating the privacy-utility trade-off of these DP methods as in Figure 5.1c is still possible. Therefore, analyzing and empirically comparing our approach with the previous work, SDP [63], is a meaningful direction for future work as well.

Chapter 7

LIMITATIONS

Although our proposed notion of DP improves model utility at the same level of privacy, SMDP, like any other DP notion, still requires improvements to preserve privacy notions as described in Brown et al. [6]. Additionally, despite our focus on language modeling tasks in this work, our novel notion of privacy can be applicable to other machine learning applications such as computer vision. Future work is required to study the effects of SMDP in other domains. It is also advised to study the fairness of SMDP since models trained with DP tend to have a disparate impact on model performance [3]. In the privatization mechanism, it is important to note that embedding models must be trained on data with a domain close to the domain of the data being privatized. The same applies to the language of embedding models and privatized data since the SMDP mechanism is dependent on the vocabulary of the embedding model. Moreover, we experiment with a simple policy function, which considers all digits as sensitive information. In real-world applications, policy functions should incorporate contextual information such as digits in the context of social security numbers. Finally, the canary insertion attack used in our experiments quantifies privacy guarantees. However, as suggested by Carlini et al. [8], these types of privacy attacks are not well-suited for real-world scenarios since models like ChatGPT output tokens instead of perplexity scores.

BIBLIOGRAPHY

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 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, 2016. Association for Computing Machinery.

[2] Samuel Adams, David Melanson, and Martine De Cock. Private text classification with convolutional neural networks. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 53–58, 2021.

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Hanna Bozakov. ChatGPT – Privacy nightmare or helpful tool?, 2023.

[6] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA, 2022. Association for Computing Machinery.

[7] 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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[8] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th*

USENIX Security Symposium (USENIX Security 19), pages 267–284, Santa Clara, CA, August 2019. USENIX Association.

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

[10] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In Emiliano De Cristofaro and Matthew Wright, editors, *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[11] Kenda Clark. As European data authorities scrutinize ChatGPT, experts see AI regulation on the horizon, 2023.

[12] Ben Derico. ChatGPT bug leaked users' conversation histories. BBC News, San Francisco, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

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

[15] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 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, 2006.

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

[18] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[19] Adel Elmahdy, Huseyin A. Inan, and Robert Sim. Privacy leakage in text classification a data extraction approach. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 13–20, Seattle, United States, July 2022. Association for Computational Linguistics.

[20] Qi Feng, Debiao He, Zhe Liu, Huaqun Wang, and Kim-Kwang Raymond Choo. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721, 2020.

[21] Natasha Fernandes, Mark Dras, and Annabelle McIver. Author obfuscation using generalised differential privacy. *CoRR*, abs/1805.08866, 2018.

[22] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing. In Flemming Nielson and David Sands, editors, *Principles of Security and Trust*, pages 123–148, Cham, 2019. Springer International Publishing.

[23] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 178–186, New York, NY, USA, 2020. Association for Computing Machinery.

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

[25] Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. Iron: Private inference on transformers. *Advances in Neural Information Processing Systems*, 35:15718–15731, 2022.

[26] Andrew Hoang, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. Efficient adaptation of pretrained transformers for abstractive summarization. *CoRR*, abs/1906.00138, 2019.

[27] Shlomo Hoory, Amir Feder, Avichai Tendler, Alon Cohen, Sofia Erell, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. Learning and evaluating a differentially private pre-trained language model. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 21–29, Online, June 2021. Association for Computational Linguistics.

[28] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[29] Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. Differentially private natural language models: Recent advances and future directions, 2023.

[30] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

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

[32] Timour Igamberdiev and Ivan Habernal. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada, July 2023. Association for Computational Linguistics.

[33] Huseyin A. Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. Privacy analysis in language models via training data leakage report. *CoRR*, abs/2101.05405, 2021.

[34] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. Membership inference attack susceptibility of clinical language models. *CoRR*, abs/2104.08305, 2021.

[35] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[36] Michael Kan. OpenAI: sorry, ChatGPT bug leaked payment info to other users. PC Magazine, Mar 24, 2023.

[37] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[38] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *CoRR*, abs/2110.05679, 2021.

[39] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[41] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[42] Lewis Maddison. Samsung workers made a major error by using ChatGPT. TechRadar, 2023.

[43] Shiona McCallum. ChatGPT banned in Italy over privacy concerns. BBC, 2023.

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

[45] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *CoRR*, abs/1609.07843, 2016.

[46] Rachel Metz. OpenAI unveils ChatGPT for businesses, stepping up revenue push. Seattle Times, Aug 28, 2023.

[47] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3), apr 2021.

[48] Takao Murakami and Yusuke Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1877–1894, USA, 2019. USENIX Association.

[49] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023.

[50] Marissa Newman. ChatGPT poised to expose corporate secrets, cyber firm warns. Bloomberg, 2023.

[51] Seong Joon Oh, Bernt Schiele, and Mario Fritz. *Towards Reverse-Engineering Black-Box Neural Networks*, pages 121–144. Springer International Publishing, Cham, 2019.

[52] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private SGD. *CoRR*, abs/1908.07643, 2019.

[53] Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[54] Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1488–1497, 2021.

[55] Katyanna Quach. Canada sticks a privacy probe into OpenAI's ChatGPT. The Register, 2023.

[56] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[58] Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, et al. Privacy-preserving classification of personal text messages with secure multi-party computation. In *Advances in Neural Information Processing Systems*, pages 3752–3764, 2019.

[59] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *CoRR*, abs/2007.07646, 2020.

[60] Michael Schade. How your data is used to improve model performance, 2023. accessed on Nov 26, 2023.

[61] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[62] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[63] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States, July 2022. Association for Computational Linguistics.

[64] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[65] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA, 2020. Association for Computing Machinery.

[66] Jingye Tang, Tianqing Zhu, Ping Xiong, Yu Wang, and Wei Ren. Privacy and utility trade-off for textual analysis via calibrated multivariate perturbations. In *Network and System Security: 14th International Conference, NSS 2020, Melbourne, VIC, Australia, November 25–27, 2020, Proceedings 14*, pages 342–353. Springer, 2020.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[68] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52, 2018.

[69] Xinwei Wu, Li Gong, and Deyi Xiong. Adaptive differential privacy for language model training. In *Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022)*, pages 21–26, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[70] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato,

Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[71] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. A differentially private text perturbation method using regularized mahalanobis metric. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 7–17, Online, November 2020. Association for Computational Linguistics.

[72] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: A review. *Artif. Intell. Rev.*, 53(6):4335–4385, aug 2020.

[73] 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. Differentially private fine-tuning of language models. *CoRR*, abs/2110.06500, 2021.

[74] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. Differential privacy for text analytics via natural text sanitization. In *Findings, ACL-IJCNLP 2021*, 2021.

Appendix A

SUPPLEMENTARY MATERIALS

Mechanisms	ε	**Original samples**: My ID is 145572. The game began development in 2010 On the morning of February 8 , 1861
MDP	6	All-New HONOUR PRCA Apple-picking. PRCA ALCS KIAA SALUTE PRCA HONOUR HONOUR TLW PRCA SBL HONOUR All-New , CityPASS
	12	Oldest Ordering The 7/29/07. The Relics gathering Northern Historical 2011 History The dinner History Anniversary Aboard , 1841
	17	Past MOB The 777677. The game began invaluable History 2011 On The evening The October Eight , 1861
SMDP	6	My ID is Apple-picking. The game began development in HONOUR On the morning of February All-New , CityPASS
	12	My ID is 7/29/07. The game began development in 2011 On the morning of February Aboard , 1841
	17	My ID is 777677. The game began development in 2011 On the morning of February Eight , 1861

Table A.1: Samples from the WikiText-2 dataset: privatized text using mechanisms at different privacy levels.

ε	1	3	6	12	17	23	29	35	41	47	52	∞
No DP	-	-	-	-	-	-	-	-	-	-	-	**22.59**
MDP	**28.89**	28.95	28.96	28.62	28.41	27.92	27.12	24.87	23.65	23.23	23.14	-
SMDP	28.90	**28.90**	**28.91**	**27.89**	**27.19**	**27.16**	**25.07**	**24.23**	**23.33**	**22.95**	**22.93**	-

Table A.2: WikiText-2 perplexity on the test set (the lower, the better). ε values were drawn from [23].

ε	1	3	6	12	17	23	29	35	41	47	52	∞
No DP	-	-	-	-	-	-	-	-	-	-	-	**4.65**
MDP	**2.68**	2.69	**2.68**	**2.70**	**2.72**	**2.70**	2.72	2.76	2.92	3.04	3.11	-
SMDP	2.69	**2.68**	**2.68**	2.71	**2.72**	2.73	**2.69**	**2.73**	**2.80**	**2.98**	**3.03**	-

Table A.3: WikiText-2 canary exposure (the lower, the better). ε values were drawn from [23].

ε	6	12	17	23	29	35	41	47	52	∞
No DP	-	-	-	-	-	-	-	-	-	**15.61**
MDP	29.23	29.19	29.20	28.56	27.33	25.23	23.37	22.93	22.11	-
SMDP	**24.01**	**23.34**	**21.91**	**21.01**	**20.46**	**20.19**	**20.03**	**19.99**	**19.90**	-

Table A.4: WikiText-103 perplexity on the test set (the lower, the better). ε values were drawn from [23].

ε	6	12	17	23	29	35	41	47	52	∞
No DP	-	-	-	-	-	-	-	-	-	**3.36**
MDP	2.65	2.68	2.68	2.51	2.56	2.52	2.65	2.41	2.52	-
SMDP	**1.78**	**2.11**	**2.43**	**2.23**	**2.22**	**2.45**	**2.42**	**2.34**	**2.39**	-

Table A.5: WikiText-103 canary exposure (the lower, the better).