Spend Your Budget Wisely: Towards an Intelligent Distribution of the Privacy Budget in Differentially Private Text Rewriting

Stephen Meisenbacher
stephen.meisenbacher@tum.de
Technical University of Munich
School of Computation, Information and Technology
Garching, Germany

Chaeeun Joy Lee
chaeeun.joy.lee@tum.de
Technical University of Munich
School of Computation, Information and Technology
Garching, Germany

Florian Matthes
matthes@tum.de
Technical University of Munich
School of Computation, Information and Technology
Garching, Germany

Abstract

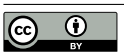
The task of *Differentially Private Text Rewriting* is a class of text privatization techniques in which (sensitive) input textual documents are *rewritten* under Differential Privacy (DP) guarantees. The motivation behind such methods is to hide both explicit and implicit identifiers that could be contained in text, while still retaining the semantic meaning of the original text, thus preserving utility. Recent years have seen an uptick in research output in this field, offering a diverse array of word-, sentence-, and document-level DP rewriting methods. Common to these methods is the selection of a privacy budget (i.e., the ε parameter), which governs the degree to which a text is privatized. One major limitation of previous works, stemming directly from the unique structure of language itself, is the lack of consideration of *where* the privacy budget should be allocated, as not all aspects of language, and therefore text, are equally sensitive or personal. In this work, we are the first to address this shortcoming, asking the question of how a given privacy budget can be intelligently and sensibly distributed amongst a target document. We construct and evaluate a toolkit of linguistics- and NLP-based methods used to allocate a privacy budget to constituent tokens in a text document. In a series of privacy and utility experiments, we empirically demonstrate that given the same privacy budget, intelligent distribution leads to higher privacy levels and more positive trade-offs than a naive distribution of ε. Our work highlights the intricacies of text privatization with DP, and furthermore, it calls for further work on finding more efficient ways to maximize the privatization benefits offered by DP in text rewriting.

CCS Concepts

• **Security and privacy → Data anonymization and sanitization**; • **Computing methodologies → Natural language processing**.

Keywords

Differential Privacy, Text Privatization, Private Text Rewriting, Data Privacy, Natural Language Processing



CODASPY '25, June 4–6, 2025, Pittsburgh, PA, USA.

ACM ISBN 979-8-4007-1476-4/25/6
https://doi.org/10.1145/3714393.3726504

ACM Reference Format:
Stephen Meisenbacher, Chaeeun Joy Lee, and Florian Matthes. 2025. Spend Your Budget Wisely: Towards an Intelligent Distribution of the Privacy Budget in Differentially Private Text Rewriting. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy (CODASPY '25), June 4–6, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3714393.3726504

1 Introduction

Efforts to address privacy preservation in Natural Language Processing (NLP) have increased in recent years, notably in the light of rapid advancements in highly advanced AI systems, primarily with Large Language Models (LLMs) [38, 48]. Such systems have enabled and fostered a nearly unfathomable array of applications for AI and NLP, yet the success of such AI systems is largely contingent upon the large-scale utilization of text data taken from a multitude of data sources, especially the Internet [15]. As such, concerns of privacy risks have continued to grow, only exacerbated by the seeming correlation between data usage and model performance [3, 37].

As a response to privacy concerns and vulnerabilities in NLP models, the field of privacy-preserving NLP (PPNLP) has steadily grown in research attention, with methods looking to bolster privacy on the data-, model-, and system-level of NLP applications [29, 41]. A popular choice among researchers in PPNLP is the framework of Differential Privacy (DP), which although was not originally intended for the unstructured domain of text [26], has seen a number of promising implementations in the literature [20].

Despite the promise, integrating DP into NLP techniques is not simple, and recent literature has unveiled a number of challenges, ranging from loss of semantics and grammatical correctness to difficulties in evaluating privacy preservation in textual data or language models [1, 26, 31]. Beyond this, other works have critiqued the manner in which DP NLP is performed, most notably relaxations in the notion of who is being protected, or what DP guarantee can be provided [35, 43]. Looking specifically to the task of *differentially private text rewriting*, in which input texts are rewritten with DP guarantees, it becomes crucial to define on which (syntactic) level a text is rewritten, and what the corresponding guarantee is [43].

Considering a text *document* as the target quantity to be rewritten, the literature is divided on how a private document can be achieved with DP rewriting. Early techniques considered the *word* to be the unit of privatization, and single-word perturbations could be composed to achieve a document-level guarantee [4, 6, 10, 12, 33, 50]. Later works also looked at the sentence level [32], or more conveniently, rewriting an entire document with one DP mechanism

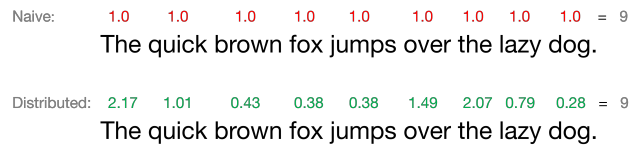


Figure 1: An example of equal privacy budget distribution vs. our proposed approach. Given that certain words may be more sensitive or revealing than others, we propose a more informed distribution (*Distributed*). This example showcases actual output from our proposed toolkit, providing a more sensible budget allocation than equal distribution.

run [22, 42]. While the benefits and limitations of these distinct approaches can be discussed, one unifying limitation is the lack of reasoning in *how* the DP privacy budget (governed by the privacy parameter ε) should be distributed to achieve a privatized document. For example, given word-level DP and a fixed document budget, the naive way may be to divide the overall budget evenly into each word, but this is certainly not optimal (see Figure 1).

In this work, we propose a more sensible method for DP text rewriting based on one simple thesis: *Not all parts of a document are equally private, and therefore, not all parts of a document should be privatized equally.* Resulting from this, we argue that a method is needed to determine a more intelligent and informed distribution of the privacy budget to a text to be rewritten. Using a toolkit of various linguistics- and NLP-based techniques, we craft a method to distribute a privacy budget sensibly for DP text rewriting, and subsequently, we leverage compositionality to achieve a final privatized text which fits into the constraints of the budget. In doing so, we answer the following research question in this work:

> *How can one intelligently "distribute" a given privacy budget in differentially private text rewriting, and what is the resulting effect on the utility and privacy of the privatized data?*

To test the utility- and privacy-preservation of our method, we compare the downstream task performance and resistance to adversarial attacks of privatized data using our distribution method to data which is *naively* privatized. We find that distributing the privacy budget with our proposed toolkit generally increases the privacy of DP rewritten text, while also leading to better trade-offs in certain cases. On the other hand, privacy budget distribution nearly always leads to lower utility and lower text coherence, leading us to critically analyze the merits and limitations of our toolkit.

As a result of our work and based on our empirical findings, we make the following contributions to the DP NLP field:

(1) We are the first to consider the distribution of privacy budget for DP text rewriting, and we propose a toolkit to determine sensible budget allocations given an input text. The toolkit is available at https://github.com/sjmeis/EpsilonDistributor.
(2) We evaluate our method in a series of privacy and utility evaluations, showing the effectiveness of budget distribution in privacy preservation.
(3) We critically analyze and discuss the implications of intelligent budget distribution for DP text rewriting, proposing ways forward to build upon our work.

2 Foundations

2.1 Differential Privacy

Formalized nearly two decades ago, Differential Privacy (DP) [9] guarantees that any computation performed on a database, or a more general collection of databases, is nearly the same regardless of the inclusion or exclusion of a single data point. Formally, given two databases $\mathcal{D}$ and $\mathcal{D}'$ differing in only one data point, any query or computation run on $\mathcal{D}$ and $\mathcal{D}'$ will yield *similar* results when utilizing some DP mechanism $\mathcal{M}$. Such databases that differ only by a single element are called *neighboring* or *adjacent* databases.

Definition 2.1. (ε, δ)-Differential Privacy. A mechanism $\mathcal{M} : \mathcal{X}^m \rightarrow \mathcal{O}$ operating on any two adjacent databases $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^m$ is (ε, δ)-differentially private, iff $\forall O \subseteq \mathcal{O}$, the following holds:

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in O] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(\mathcal{D}') \in O] + \delta$$

where $\varepsilon > 0$ and $\delta \in [0, 1]$

Intuitively, ensuring the above privacy guarantee grants plausible deniability to individuals participating in a database, such that the result of some query cannot be attributed to this person's participation in a database. Instead, the DP mechanism $\mathcal{M}$ grants this deniability, usually achieved by the injection of calibrated random noise to queries or computations.

In our work, we utilize the DP-BART rewriting mechanism [22], which guarantees (ε, δ)-DP for any two *documents*.

2.2 Metric Differential Privacy

In some domains, such as that of natural language and textual data, the original notion (ε, δ)-DP may be too restrictive, or rather not fitting to the reasoning of the "individual" in a dataset. As such, the notion of *Metric* Differential Privacy (MDP) has emerged in recent years to address the limitation [5]. It is most useful when dealing in *metric spaces*, and it can be defined as follows.

Let $\mathcal{X}$ and $\mathcal{Z}$ be finite sets and let $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}+$ be a distance metric defined on the set $\mathcal{X}$.

Definition 2.2. ($d_{\mathcal{X}}$-privacy). Let $\varepsilon > 0$. A randomized mechanism $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Z}$ satisfies $\varepsilon d_{\mathcal{X}}$-privacy iff $\forall x, x' \in \mathcal{X}$ and $\forall z \in \mathcal{Z}$

$$\frac{\mathbb{P}[\mathcal{M}(x) = z]}{\mathbb{P}[\mathcal{M}(x') = z]} \leq e^{\varepsilon d(x,x')} \tag{1}$$

The above can clearly be seen as a relaxation of Definition 2.1, as the privacy guarantee is now scaled according to the distance between any two data points in a given space. Intuitively, when an MDP mechanism is applied, queries on data points which are close in space, as measured by the chosen metric, would yield more "similar" output distributions as compared to points farther apart.

In this work, we utilize the 1-Diffractor mechanism [33], which leverages word-level MDP to provide guarantees for any two *words*.

2.3 Local DP and Text Rewriting

As opposed to the setup where data is first collected by some aggregator before applying a chosen DP mechanism, known as *Global* Differential Privacy, the concept of *Local* Differential Privacy (LDP) becomes useful in cases where third party aggregators are not trusted or where privatization must occur on the user side. In LDP,

the notion of adjacent databases is shifted to the individual, and it is defined over data points from a single individual. Thus, every collected data point from a single individual is adjacent to every other data point from another individual [24]. Note that LDP is also defined for MDP, thus yielding MLDP [36].

In the case of DP text rewriting, the LDP setup is most sensible, so that users can privatize their text(s) via rewriting before sharing it with third parties. In this scenario, the user utilizes a DP mechanism operating on a particular syntactic level, and they privatize their textual data accordingly. For example, in a word-level scenario, a user shares each obfuscated word, whereby documents can be shared according to the composition theorem of DP (see next). Similarly, if the users opts to use a document-level mechanism, the output of each mechanism run is a privatized document with an accompanying privacy guarantee. As noted by Vu et al. [43], the distinction of granularity is particularly crucial in the case of text privatization with DP, as it must be made transparent for which syntactic unit a guarantee is being provided.

The limitation with the LDP setup, however, is that for the given unit of protection (e.g., a word or document), any data point from one user is adjacent to the entire space of data points. For example, in the case of documents, any potential text document is adjacent to any other document. This limitation is highlighted by Igamberdiev and Habernal [22], which above all necessitates higher privacy budgets for sensible privatization.

2.4 Composition in DP Text Rewriting

When reporting privacy guarantees in DP text rewriting scenarios, it becomes very important to leverage the composition theorem of DP, which is defined as follows:

Theorem 2.3. *Composition in DP [9].*
Let M_1 be an ε_1-differentially private algorithm, and let M_2 be an ε_2-differentially private algorithm. Then their combination, defined to be $M_{1,2}$: $M_{1,2}(x) = (M_1(x), M_2(x))$, is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

The implications of the above theorem are quite useful in reporting aggregate privacy guarantees: if one runs a DP mechanism with privacy budget ε for n times, then the resulting guarantee is $n \cdot \varepsilon$. The intuition is also clear; privacy guarantees begin to degrade the more times a mechanism is used on the same data.

In DP text rewriting, composition can be leveraged to utilize DP mechanisms a number of times to achieve the desired syntactic unit of privatization. For example, given a text of 10 words, a word-level DP mechanism can be run on each of the 10 words with a budget of ε per word for a total guarantee of 10ε. This can naturally be extrapolated to sentences in a document, documents per user, and beyond. For the purposes of this work, we treat the document as the final unit of protection, although recent work has shown that this assumption may not always be correct [43].

In this work, we place a particular focus on the question of composition in DP text rewriting, investigating whether this theorem can be leveraged more wisely by considering the hierarchical nature of textual data. Specifically, we consider the scenario where a fixed privacy budget is allotted for each document to be privatized, and we explore how this budget can be maximized to protect the privacy concealed in natural language, while still maintaining the utility of text datasets. We challenge the "naive" distribution of privacy budgets, in which, for example, a document is privatized singly without higher focus on more sensitive sentences, or similarly where a document is privatized with equal emphasis on all words rather than an intelligent distribution of stricter privatization to more sensitive words. To address this, we now introduce a toolkit of techniques that will allow for a more informed and sensible privacy budget distribution in DP text rewriting.

3 A Budget Distribution Toolkit for DP Text Rewriting

In this section, we introduce the underlying methodology behind the distribution of a given privacy budget over a text document. The goal of such a distribution is to allocate a privacy budget intelligently so as to account for a number of linguistic- and NLP-informed factors which may make certain tokens in a document more sensitive than others. We first outline the general framework for budget distribution, and then we proceed to introduce the individual components of our proposed toolkit.

3.1 Allocating ε

We consider an arbitrary text document $\mathcal{D}$ to be privatized via DP text rewriting. The document $\mathcal{D}$ consists of n tokens, or words, which are sequential in nature, i.e., $D = (t_i)_n^1$, where t_i denotes a token in the i-th position in the document string.

We also define a general scoring function $\mathcal{S}(\mathcal{D}, \varepsilon)$, which takes as input an arbitrary document $\mathcal{D}$ and a privacy budget ε. The output of $\mathcal{S}$ is a mapping $\mathcal{S} : t_i \rightarrow \mathbb{R}^+ = s_i, \forall i \in \{1..n\}$. Thus, each token in the document $\mathcal{D}$ is assigned a normalized *sensitivity score* s, where a higher score denotes a greater "need" for privatization. In the context of DP and the total privacy budget ε, this translates to the allocation of a smaller token budget. Formally, we are therefore solving the linear equation $Ax = b$, where

$$A = \begin{bmatrix} \frac{1}{s_1} & \frac{1}{s_2} & \dots & \frac{1}{s_{n-1}} & \frac{1}{s_n} \end{bmatrix}, b = \varepsilon$$

Thus, the resulting solution for the equation $\mathbf{x}$ is a 1:1 mapping of constituent tokens to per-token budget allocations:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_{n-1} & x_n \end{bmatrix} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_{n-1} & \varepsilon_n \end{bmatrix}$$

And finally, by leveraging compositionality, we can achieve a distribution that respects the original total privacy budget ε:

$$\sum_{i=1}^{n} \varepsilon_i = \varepsilon$$

Note that the case of per-token privacy budget allocations can be generalized to the sentence level in document privatization by simply summing scores of the constituent tokens of a sentence, thus yielding a sentence-level privacy budget allocation.

3.2 Budget Allocation Methods

Our toolkit consists of five methods used to calculate per-token budget allocations, which are introduced below, as well as the technique used to combine the component scores into a final allocation. For full method details, we refer to reader to our code repository.

*3.2.1 Information Content (**IC**).* Information Content (IC), also referred to as *self-information* or *Shannon information*, is a value derived from measuring the probability of a particular event occurring. In the context of linguistics, such a value can be assigned to any given unit in language, most notably a word token.

In leveraging IC measures, we base the resulting scores on the hypothesis that the greater the relative information given by a particular word is, the greater the need for privacy. We utilize the IC measures provided by the NLTK packages, namely the *semcor*, *brown*, *bnc*, *shaks*, and *treebank* corpora. We also use the English WordNet, which contains *synsets*; these synsets, or word entries, are required as the input format for retrieving the IC scores.

As the abovementioned corpora only assign IC scores to nouns and verbs, we likewise only score nouns and verbs; these IC scores are given on the positive integer range (e.g., *IC*('dog') = 235). Non-noun/verb tokens, as well as tokens not existing in the corpora, are assigned a score of 1, or the lowest possible positive integer.

*3.2.2 Part-of-Speech Informativeness (**POS**).* Continuing with our hypothesis that word informativeness can help to determine the level of privacy needed, we also leverage Part-of-Speech (POS) information to calculate informativeness scores. POS tags indicate the grammatical function of a word (noun, verb, adjective, etc.) in a text. Therefore, assigning different weights to different POS tags can reflect their typical importance in conveying meaning. For example, verbs often play a more central role in a sentence compared to prepositions. This helps distinguish between function words (like articles and prepositions) and content words (nouns, verbs, adjectives) that contribute more to the core meaning [25].

To define a weighting scheme for different POS tags, we refer to a previous study based on Twitter data [14], and use the aggregate statistics to derive weights that denote the relative frequency of POS tags. We focus on *nouns* (NN), *pronouns* (PR), *verbs* (VB), *adjectives* (JJ), *adverbs* (RB), and *numbers* (CD), with the following weights: {NN: 14, PR:7, VB:15, CD:2, JJ:5, RB:5}. All POS tags outside this set are not considered sensitive, and they receive a weight of 0.1, chosen to be distinct from the abovementioned values yet to assign non-zero weights to avoid division errors.

*3.2.3 Named Entity Recognition (**NER**).* The task of Named Entity Recognition (NER) aims to identify *named entities* in a given text, such as names, locations, and organizations. These entities are typically very important to a particular sentence's meaning; however, they are generally quite identifying, such as with names.

We use the NER tool provided by the SPACY package[1] to identify named entities in a given input document, and all tokens that belong to a named entity are assigned a score of 1, otherwise a score of 0.

*3.2.4 Word Importance (**WI**).* This method compares the semantic similarity between the entire text document and each individual word. Words with a larger difference in similarity likely contribute more to the overall meaning, as they introduce new semantic meaning not already conveyed by the rest of the text.

To measure such importance of each word, we iteratively remove each word token from a given text, and measure its similarity to the remainder of the text. The lower the similarity between these two entities, the greater the importance of the word.

[1] https://spacy.io/

*3.2.5 Sentence Difference (**SD**).* In a similar way to the above *Word Importance* scoring, we also measure the semantic difference between a given text and the same text without a single word. This, similar to the above, provides a notion of a word's importance semantically. Words whose removal causes a more significant drop in similarity are considered more important because their absence significantly impacts the overall meaning. Thus, such words are more identifiable in text and must be treated with higher privacy.

To measure *Sentence Difference*, we create n versions of the original sentence, each with one of $t_1 ... t_n$ removed. These versions are then compared semantically to the original, unaltered sentence, and the resulting scores are assigned to corresponding tokens.

For both the **WI** and **SD** methods, we utilize the THENLPER/GTE-SMALL embedding model[2] [28].

3.3 Calculating a Final Budget Distribution

Given the scores outputted by each of the five methods described above, the final steps involve combining these scores for a final privacy budget distribution for DP text rewriting.

First, all score sets, which map a score to each token in an input document, are normalized between 0 and 1. This is particularly necessary in the case of **POS**, where the weights assigned do not fall in the range [0,1]. Then, the average score of each token amongst the five scoring methods is taken to achieve an aggregate score for each token. Note that we assume an equal weighting for each of the methods, and we do not experiment with different weighted averages. However, please refer to Section 4.4 and Table 4 for the results of our ablation study for the described methods.

Given the aggregate scores for all word tokens in a document, the linear equation as described in Section 3.1 is solved. This results in an individual budget for each token, all of which add up to the total allocated privacy budget ε.

4 Experimental Setup and Results

Following the guidelines of Mattern et al. [31] of what comprises an effective text privatization method, we evaluate the performance of our proposed budget distribution method on two primary categories: privacy protection and utility preservation. These evaluations will demonstrate to what degree our method improves upon the empirical privacy protections afforded by DP text rewriting, while simultaneously testing whether utility in downstream tasks and text coherence can still be achieved. In the following, we outline the full methodological design of our experiments, as well as provide the corresponding results.

4.1 DP Text Rewriting Methods

In the scope of this work, we choose two DP text rewriting methods from the recent literature, which will serve as the testbed for our proposed budget distribution toolkit.

1-DIFFRACTOR [33]. 1-DIFFRACTOR is a word-level MLDP text obfuscation mechanism proposed by Meisenbacher et al. to improve the efficiency of previous word-level mechanisms. In essence, the mechanism *perturbs* words in a document by adding DP noise to word embeddings in one-dimensionally sorted lists. In this work,

[2] https://huggingface.co/thenlper/gte-small

we use the *geometric* version of the 1-DIFFRACTOR mechanism, or $1 - D_G$. Following from the original work, we test the following base, per-word ε values: $\varepsilon \in \{0.1, 0.5, 1.0\}$.

DP-BART [22]. DP-BART is a DP text rewriting mechanism proposed by Igamberdiev and Habernal which achieves DP rewriting at the document level by adding calibrated DP noise in the latent space representation of the BART encoder-decoder model [27]. In this work, we utilize the DP-BART-CLV version, which achieves DP by clipping the latent space before adding noise. Following from the original work, we choose the clipping range of $[-0.1, 0.1]$, and we test on the following document-level ε values: $\varepsilon \in \{500, 1000, 1500\}$.

4.1.1 Setting the total privacy budget. In the evaluation of both of the abovementioned rewriting methods on our chosen datasets and tasks (see next), we must first determine the total privacy budget (ε) allocated to each text document to be privatized. This is especially pertinent in the case of our chosen *word-level* mechanism, as we aim to privatize documents at the *document* level.

Following the example set in previous work [34], we fix a *dataset-specific* per-document privacy budget, which can be derived as the chosen base ε value, scaled (multiplied) by the average number of tokens in a document for a given dataset. Thus, for 1-DIFFRACTOR, our chosen values of $\varepsilon \in \{0.1, 0.5, 1.0\}$ will be scaled for each dataset to achieve the total privacy budget available for each document. The exact values used for these calculations and the per-document budgets are made apparent in all tables presenting results.

Note that for DP-BART, the chosen ε values already represent the per-document budgets, as this mechanism operates directly on the document level. For both 1-DIFFRACTOR and DP-BART, we test for both an equal distribution of the total privacy budget, i.e., ε/num_tokens, as well as budget distribution with our proposed toolkit. With DP-BART, the resulting per-word budgets are summed to achieve sentence-level budgets. Thus, an input document is split into its component sentences, and each sentence is privatized with DP-BART according to the allocated budget. In all cases, stopwords (common English words as determined by the NLTK package) are not considered in the budget distribution and are not privatized.

4.2 Privacy Experiments

We run experiments to evaluate the privacy-preserving capability of our method relative to naive (equal) budget distribution. These experiments take two forms: empirical privacy and membership inference. We first introduce the datasets used for experimentation, and then we proceed to describe in detail the evaluation procedures.

4.2.1 Datasets and Tasks. For the privacy experiments, we leverage three existing public datasets.

Yelp Reviews. We utilize a dataset of reviews from the *Yelp* platform, specifically the subset made available by Utpala et al. [42][3]. This subset contains 17,295 reviews from 10 distinct users on the platform. Each review is denoted as a *positive* or *negative* review in terms of sentiment. This dataset's makeup allows for an adversarial *authorship identification* task, in which an attacker's goal is to guess the identity of the text's author given only the text.

[3]The full dataset is available at https://huggingface.co/datasets/Yelp/yelp_review_full.

Trustpilot Reviews. Made available by Hovy et al. [18], the Trustpilot Reviews corpus is a collection of reviews in several languages from the Trustpilot platform. We only use English-language reviews from the *en-US* subset, and we take a 10% sample (29,490 reviews). Each review is marked as positive or negative, as well as with the gender (male/female) of the author, creating the opportunity for evaluation on an adversarial *gender identification* task.

Blog Corpus. The final dataset we use for our privacy experiments is a subset from the Blog Authorship Corpus [40], a large collection of user-written blog posts on an internet forum. In particular, we make use of the *author10* split made available by Meisenbacher and Matthes [35], which contains a total of 15,070 blog posts from the top-10 contributing authors. Thus, we create another authorship identification scenario for our empirical privacy experiments.

4.2.2 Empirical Privacy Evaluation. The first of two overarching privacy evaluation tasks takes the form of *empirical privacy* evaluations. Here, we test the ability of DP text rewriting to reduce the adversarial advantage (i.e., attribute inference performance) on authorship or gender identification, measured *empirically*.

To test empirical privacy, we first privatize all of the above datasets using our two chosen DP rewriting methods under the chosen privacy budgets. This is done for both naive budget distribution and distribution using our toolkit. Then, we train an adversarial classification model to predict the protected attribute (author or gender) given a text. For all experiments, a DEBERTA-V3-BASE model [17] is used, and datasets are split into a 90% train / 10% test set.

We perform the adversarial training in two settings, following the recent literature [31, 35, 42]. In the first, called the *static* setting, the adversarial classification model is trained on the non-privatized train set, and the resulting model is evaluated on the privatized test set for the *static* results. This models a less capable attacker who does not have knowledge of the DP rewriting method. In the more capable setting, called the *adaptive* attacker, the adversarial model is trained on the privatized train set, and then evaluated on the privatized test set, thus mimicking an adversary who is able to train a better model given the ability to align the training dataset. For all scenarios in these experiments and for the remainder of this work, training is performed for one full epoch, with a batch size of 32, maximum input length of 512 tokens, learning rate of 1e-5, and otherwise default HuggingFace TRAINER parameters. All training procedures are repeated three times on different random shuffles of the train set, and the report scores represent the average score with standard deviation. The hardware used is an RTX A6000 GPU.

For score reporting, we first measure the corresponding utility of each privatized dataset, measured by the micro-F1 score on the binary sentiment analysis task after one epoch of training. As the datasets are imbalanced (many positive reviews), we also provide *PP+*, or the percentage points achieved over majority-class guessing.

Next, we report the adversarial F1 score against the non-privatized (plaintext) baseline, where a lower score denotes that the DP rewriting has better protected privacy. Finally, we report the *Relative Gain (γ)* metric [31, 46], which aims to illustrate the balance between (potential) utility lost and privacy gained. Let P_o, U_o represent the baseline privacy and utility scores, respectively, and P_r, U_r be the scores observed on the privatized datasets. Relative Gain is thus

Table 1: Empirical Privacy Results. Adversarial scores are represented by micro-F1 scores, where *(s)* denotes the *static* attacker and *(a)* denotes the *adaptive* attacker. γ refers to the *Relative Gain* for both the static and adaptive settings. For all 1-DIFFRACTOR tasks, we report the per-word ε, as well as the total allocated budget (indicated in parentheses), which is calculated by (*per-word* ε)·(*Avg. Tokens*), or average number of tokens in a document per dataset. Baseline scores, i.e., adversarial performance on the non-privatized data, are also provided. Where γ scores are reported, the bolded score represents the better pairwise score between a "naive" distribution and our proposed method.

Yelp		**1-Diffractor**						**DP-BART**					
Avg. Tokens		181.06						181.06					
ε		0.1 (18.11)		0.5 (90.53)		1 (181.06)		500		1000		1500	
Distribution Method		Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Utility (F1) ↑	$95.09_{0.3}$	$93.53_{0.0}$	$93.53_{0.0}$	$95.01_{0.6}$	$94.01_{0.9}$	$94.70_{1.0}$	$94.45_{0.6}$	$93.53_{0.0}$	$93.53_{0.0}$	$93.99_{0.1}$	$93.53_{0.0}$	$94.47_{0.7}$	$93.53_{0.0}$
PP+ ↑	-156	-312	-312	-164	-264	-195	-220	-312	-312	-266	-312	-218	-312
Adv. F1 (s) ↓	95.90	42.20	42.37	57.23	55.09	65.84	62.60	26.24	17.11	26.99	15.09	27.86	15.55
Adv. F1 (a) ↓	95.90	$80.92_{3.0}$	$82.35_{0.9}$	$92.16_{0.1}$	$88.44_{1.4}$	$92.72_{0.5}$	$92.87_{0.7}$	$38.82_{1.0}$	$32.87_{0.4}$	$61.21_{0.6}$	$37.30_{0.8}$	$67.63_{1.2}$	$38.71_{0.6}$
γ (s) ↑	-	**1.81**	**1.81**	0.64	**1.31**	0.19	**0.91**	2.06	**2.19**	1.75	**2.22**	1.43	**2.22**
γ (a) ↑	-	**1.23**	1.21	0.11	**0.81**	0.30	**0.46**	1.86	**1.95**	1.23	**1.89**	0.83	**1.87**

(a) Yelp

Trustpilot		**1-Diffractor**						**DP-BART**					
Avg. Tokens		51.23						51.23					
ε		0.1 (5.12)		0.5 (25.62)		1 (51.23)		500		1000		1500	
Distribution Method		Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Utility (F1) ↑	$99.49_{0.1}$	$94.87_{1.3}$	$93.62_{2.3}$	$98.15_{0.2}$	$97.93_{0.3}$	$98.89_{0.1}$	$98.64_{0.6}$	$92.59_{0.4}$	$92.09_{0.1}$	$98.03_{0.1}$	$92.16_{0.2}$	$98.51_{0.1}$	$93.36_{0.1}$
PP+ ↑	366	-96	-221	232	210	306	281	-324	-374	220	-367	268	-247
Adv. F1 (s) ↓	72.16	59.61	59.51	64.12	63.45	67.28	66.40	60.71	58.94	59.85	59.34	60.16	59.34
Adv. F1 (a) ↓	72.16	$60.71_{3.7}$	$60.37_{3.2}$	$67.70_{2.2}$	$66.59_{0.6}$	$63.87_{4.1}$	$68.25_{2.2}$	$58.83_{0.6}$	$58.09_{0.0}$	$62.48_{0.9}$	$58.12_{0.0}$	$61.18_{1.8}$	$58.94_{1.2}$
γ (s) ↑	-	**1.02**	0.70	1.09	**1.16**	**1.34**	0.81	0.20	**0.38**	**1.84**	0.33	**1.91**	0.66
γ (a) ↑	-	**0.82**	0.54	0.44	**0.59**	**1.34**	0.48	**0.54**	**0.54**	**1.36**	0.55	**1.73**	0.73

(b) Trustpilot

Blog		**1-Diffractor**						**DP-BART**					
Avg. Tokens		53.94						53.94					
ε		0.1 (5.39)		0.5 (26.97)		1 (53.94)		500		1000		1500	
Distribution Method		Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Adv. F1 (s) ↓	58.64	26.56	26.19	31.34	31.96	35.39	35.43	12.73	7.95	15.05	12.51	15.87	17.22
Adv. F1 (a) ↓	58.64	$40.05_{0.8}$	$38.55_{1.0}$	$44.89_{1.7}$	$44.49_{0.8}$	$43.96_{9.6}$	$46.89_{1.5}$	$13.61_{0.5}$	$8.95_{0.8}$	$21.77_{0.9}$	$13.58_{1.9}$	$23.77_{1.0}$	$16.60_{3.5}$

(c) Blog

defined as $\gamma = (U_r/U_o) - (P_r/P_o)$, with the higher the better. Different to previous work, we calculate the change in F1 over random / majority-class guessing on the validation set, denoted MG_u (utility) and MG_p (privacy), as the Trustpilot and Yelp datasets are imbalanced; thus $RG = \frac{U_r - MG_u}{U_o - MG_u} - \frac{P_r - MG_p}{P_o - MG_p}$.

In the Yelp dataset, the 10% validation split contains 1618 positive reviews and 112 negative reviews. Thus, $MG_u = 96.65$. The split contains 304 reviews from the most frequent author, with the nine other authors writing 1426 reviews. Thus, $MG_p = 29.89$, showing the majority-class guessing performance. In the Trustpilot dataset, the 10% validation split contains 2713 positive reviews and 236 negative reviews. Thus, $MG_u = 95.83$. The split contains 1713 reviews from males and 1236 reviews from females. Thus $MG_p = 66.67$. Note that we use random guessing performance due to the relative balance between male and female authors.

The complete results of the empirical privacy experiments can be found in Table 1, for both the static (s) and adaptive (a) settings. Note that as the Blog dataset does not have an associated utility task, we do not report utility or γ values.

4.2.3 Membership Inference Evaluation. In the context of privacy-preserving Machine Learning, *Membership Inference Attacks* (MIAs) attempt to infer whether a specific data record (e.g., individual) was part of the data used to train a model [19]. With textual data, MIAs take on a slightly different meaning, and essentially, the goal of the attacker becomes to infer whether certain textual information was present in the training data [38]. To evaluate the resilience of DP text rewriting against MIAs, we run two types of experiments.

Masked Token Inference Attack (MTI). Following Chen et al. [7], we run a *masked token inference attack*. We leverage the ability of masked language models (MLMs) to predict a masked (hidden) word given the context surrounding the word. Thus, given a privatized text, we test an MLM's ability to predict tokens from the original text when provided with the privatized context. To measure the performance of this attack, we follow a procedure as such:

(1) For each privatized document, mask each token one by one. In this work, we use BERT-BASE-UNCASED [8].
(2) Capture the top predictions of the MLM (top-1 and top-3).

Table 2: Membership Inference Evalaution. Bolded scores represent the better score between "naive" and our proposed distribution (shown only for MTI_{bow} and NN). For further details on each metric, please refer to Section 4.2.3.

			$MTI_{seq}T1 \downarrow$	$MTI_{seq}T3 \downarrow$	$MTI_{bow}T1 \downarrow$	$MTI_{bow}T3 \downarrow$	NN↑
Yelp	0.1	Naive	0.003	0.007	**0.121**	**0.151**	111
		Ours	0.003	0.007	**0.121**	**0.151**	**154**
	0.5	Naive	0.003	0.006	0.122	**0.153**	9
		Ours	0.003	0.006	**0.120**	**0.153**	**17**
	1	Naive	0.003	0.006	0.124	0.155	2
		Ours	0.003	0.006	**0.123**	**0.154**	**4**
Trustpilot	0.1	Naive	0.002	0.004	**0.055**	**0.074**	308
		Ours	0.002	0.004	**0.055**	0.075	**383**
	0.5	Naive	0.002	0.004	0.059	0.078	28
		Ours	0.002	0.004	**0.057**	**0.077**	**64**
	1	Naive	0.002	0.004	0.062	0.081	5
		Ours	0.002	0.003	**0.060**	**0.079**	**15**
Blog	0.1	Naive	0.001	0.002	**0.052**	**0.067**	47
		Ours	0.001	0.002	**0.052**	**0.067**	**69**
	0.5	Naive	0.001	0.002	**0.052**	**0.068**	3
		Ours	0.001	0.002	0.053	**0.068**	**6**
	1	Naive	0.001	0.002	0.053	**0.068**	2
		Ours	0.001	0.001	**0.052**	**0.068**	**3**

(a) 1-DIFFRACTOR

			$MTI_{seq}T1 \downarrow$	$MTI_{seq}T3 \downarrow$	$MTI_{bow}T1 \downarrow$	$MTI_{bow}T1 \downarrow$	NN↑
Yelp	500	Naive	0.002	0.005	0.104	0.137	816
		Ours	0.001	0.004	**0.042**	**0.135**	**964**
	1000	Naive	0.002	0.004	0.109	0.135	342
		Ours	0.001	0.004	**0.043**	**0.079**	**936**
	1500	Naive	0.002	0.005	0.112	0.139	203
		Ours	0.001	0.003	**0.051**	**0.087**	**879**
Trustpilot	500	Naive	0.002	0.004	0.052	0.068	780
		Ours	0.001	0.005	**0.019**	**0.035**	**956**
	1000	Naive	0.001	0.003	0.058	0.074	257
		Ours	0.002	0.004	**0.025**	**0.041**	**787**
	1500	Naive	0.001	0.002	0.061	0.076	146
		Ours	0.001	0.002	**0.032**	**0.049**	**604**
Blog	500	Naive	0.001	0.002	0.042	0.058	772
		Ours	0.001	0.002	**0.024**	**0.036**	**833**
	1000	Naive	0.001	0.001	0.044	0.056	520
		Ours	0.001	0.001	**0.026**	**0.041**	**559**
	1500	Naive	0.001	0.002	0.049	0.061	**472**
		Ours	0.001	0.001	**0.029**	**0.045**	406

(b) DP-BART

(3) Check if the predictions match the exact original token in sequence (MTI_{seq}), or if the predictions match any token in the original text (MTI_{bow}), as in a bag-of-words.

Thus, for each dataset, we report four scores: $MTI_{seq}T1$, $MTI_{seq}T3$, $MTI_{bow}T1$, and $MTI_{bow}T3$, where T1 and T3 represent considering the top-1 and top-3 predictions, respectively. For all scores, a lower score means higher privacy protection.

Nearest Neighbor Attack (NN). We also design a new attack, called the *nearest neighbor attack*, which measures how close (semantically), on average, the privatized text is to the original text given the entire privatized dataset. The procedure is as follows:

(1) For each document in the original dataset, select this document as the *query* document.
(2) Note the index of the query's private counterpart in the privatized dataset, or the *corpus*.
(3) Using an embedding model and cosine similarity measures, measure for which k value the private document is the k-th nearest neighbor to the query document.

With this, we measure the *plausible deniability* that is created, i.e., the "distance" from the original document to the private document. We report the average k over all documents in the private dataset. In the context of MIA, a higher average k would imply higher privacy.

The results of both the MTI and NN experiments on all three privacy datasets can be found in Table 2.

4.3 Utility Experiments

In addition to measuring the privacy-preserving capabilities of the DP rewriting methods both with and without budget distribution, we also measure the utility preservation, namely the effect on downstream task utility between a naive and our proposed distribution.

4.3.1 Datasets and Tasks. We use six datasets for utility evaluation.

GLUE Datasets. The GLUE Benchmark [44] consists of nine datasets focused on evaluating the general language understanding capabilities of language models. We choose three datasets from each of the three sub-tasks of the benchmark: SST-2 (sentiment analysis), MRPC (sentence similarity), and MNLI (textual entailment). In the case of the MNLI, we take a 10% subset for a total of 39,270 training instances. Note that since these datasets are all a maximum of one sentence, we only evaluate them with 1-DIFFRACTOR.

BBC News. The BBC News dataset[4] is a collection of 3147 news articles from the BBC platform, where each news article belongs to one of five popular news categories: *business*, *entertainment*, *politics*, *sports*, or *tech*. This creates a five-class classification task.

DocNLI. The *DocNLI* dataset [49] introduces a *document*-level entailment prediction task. The dataset consists of (*premise*, *hypothesis*) pairs, each marked as *entailment* or *not entailment*. As the original dataset is very large, we take a 1% random sample, resulting in a 9136-row dataset with two classes.

IMDb Reviews. The IMDb Dataset [30] consists of 50k movie reviews from the IMDb platform, each labeled as positive or negative.

4.3.2 Utility Evaluation. To evaluate utility, we follow a similar procedure to the model training described in Section 4.2.2. Firstly, each dataset is privatized using our two chosen DP rewriting methods, their respective three ε values, and the two distribution techniques. The resulting datasets, including the original baseline datasets, are used to train a DEBERTA-V3-BASE classification model for one epoch.

For each training procedure, we report the F1 score achieved by the trained model on the 10% held-out test set. These results are found in Table 3. In addition, we report three other metrics to capture the quality and coherence of the privatized text documents:

(1) *Cosine Similarity* (CS): we measure the average cosine similarity of the embeddings of the original texts and their privatized counterparts, using Sentence Transformers [39], specifically with the ALL-MINILM-L12-V2 embedding model.
(2) *BLEU*: the bilingual evaluation understudy (BLEU) score is used to measure the quality of generated text (i.e., private texts) as compared to a reference text (i.e., original texts). We report the average BLEU score using the NLTK package.
(3) *Perplexity* (PPL): *perplexity* can be used as a proxy to measure the coherence and understandability of a text, as it measures how "surprised" a language model is when seeing a given

[4] http://mlg.ucd.ie/datasets/bbc.html

Table 3: Utility Experiment Results. Utility scores are represented by F1 scores achieved on the corresponding tasks. For all 1-DIFFRACTOR tasks, we report the per-word ε, as well as the total allocated budget (indicated in parentheses), which is calculated by (*per-word* ε)·(*Avg. Tokens*), or average number of tokens in a document per dataset. Baseline scores on the non-privatized data are also provided. Note that for DP-BART and the *IMDb* dataset, we take a 20% random split due to the size of the dataset.

	SST2						MRPC						MNLI					
Baseline	$96.12_{0.1}$						$86.56_{0.9}$						$86.68_{0.2}$					
Avg. Tokens	8.31						18.29						19.54					
ε	0.1 (0.83)		0.5 (4.16)		1.0 (8.31)		0.1 (1.83)		0.5 (9.15)		1.0 (18.29)		0.1 (1.95)		0.5 (9.77)		1.0 (19.54)	
Distribution	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Utility (F1) ↑	$80.40_{0.3}$	$78.93_{0.2}$	$87.21_{0.1}$	$85.31_{0.2}$	$91.20_{0.1}$	$88.94_{0.1}$	$70.12_{1.4}$	$64.67_{0.1}$	$75.93_{1.1}$	$74.21_{1.8}$	$71.39_{2.3}$	$73.48_{6.6}$	$40.37_{6.6}$	$39.78_{5.7}$	$56.74_{12.8}$	$55.30_{4.5}$	$67.77_{0.9}$	$67.26_{2.0}$
CS ↑	0.508	0.459	0.707	0.625	0.813	0.721	0.438	0.382	0.665	0.573	0.786	0.689	0.464	0.406	0.706	0.610	0.819	0.719
BLEU ↑	0.163	0.155	0.240	0.203	0.341	0.267	0.126	0.118	0.205	0.174	0.314	0.243	0.164	0.157	0.261	0.220	0.372	0.297
PPL ↓	9372	9849	6329	7266	5152	5813	1808	1912	991	1233	552	795	1892	2035	1361	1232	648	857

(a) 1-Diffractor (1/2)

	BBC						DocNLI						IMDb					
Baseline	$98.73_{0.3}$						$87.82_{0.4}$						$95.84_{0.2}$					
Avg. Tokens	399.34						285.22						225.97					
ε	0.1 (39.99)		0.5 (199.97)		1.0 (399.94)		0.1 (28.52)		0.5 (142.61)		1.0 (285.22)		0.1 (22.60)		0.5 (112.99)		1.0 (225.97)	
Distribution	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Utility (F1) ↑	$92.17_{0.6}$	$75.87_{15.6}$	$95.87_{0.0}$	$95.56_{0.3}$	$96.83_{0.4}$	$95.77_{1.6}$	$79.10_{0.5}$	$78.30_{0.5}$	$82.02_{0.2}$	$80.34_{1.0}$	$82.86_{0.3}$	$82.02_{0.8}$	$57.31_{9.9}$	$77.64_{14.8}$	$92.73_{0.2}$	$92.77_{0.1}$	$94.46_{0.0}$	$93.94_{0.1}$
CS ↑	0.507	0.461	0.753	0.699	0.855	0.807	0.593	0.546	0.814	0.765	0.879	0.845	0.471	0.421	0.707	0.639	0.818	0.750
BLEU ↑	0.132	0.127	0.222	0.196	0.330	0.279	0.112	0.107	0.190	0.167	0.248	0.218	0.167	0.162	0.263	0.234	0.370	0.317
PPL ↓	692	725	324	397	173	236	716	795	311	376	212	257	621	655	309	371	186	240

(b) 1-Diffractor (2/2)

	BBC						DocNLI						IMDb*					
Baseline	$98.73_{0.3}$						$87.82_{0.4}$						$95.13_{0.2}$					
Avg. Tokens	399.34						285.22						225.97					
ε	500		1000		1500		500		1000		1500		500		1000		1500	
Distribution	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours	Naive	Ours
Utility (F1) ↑	$40.53_{1.8}$	$32.17_{1.2}$	$90.48_{0.4}$	$33.86_{1.3}$	$92.59_{0.1}$	$32.59_{0.8}$	$66.96_{2.3}$	$75.89_{0.8}$	$75.71_{0.9}$	$74.36_{2.1}$	$75.46_{0.9}$	$76.29_{0.8}$	$70.23_{1.3}$	$49.23_{1.0}$	$87.77_{0.6}$	$50.20_{1.9}$	$89.07_{0.3}$	$53.37_{2.4}$
CS ↑	0.145	0.065	0.414	0.072	0.489	0.072	0.168	0.069	0.453	0.104	0.531	0.176	0.187	0.091	0.411	0.102	0.454	0.129
BLEU ↑	0.017	0.000	0.051	0.000	0.065	0.000	0.012	0.000	0.038	0.001	0.047	0.004	0.028	0.000	0.075	0.002	0.090	0.006
PPL ↓	23	34667	11	15207	11	5811	27	18314	12	8303	11	4936	19	25054	9	5351	9	1908

(c) DP-BART

text. We use a GPT-2 model to measure perplexity [35, 46]. For performance, we limit input texts to the first 256 tokens.

4.4 Ablation Study

The final component of our experiments involves an ablation study with our privacy budget distribution toolkit, namely to measure the individual effect of the five proposed scoring methods. Thus, we are able to identify which of the methods leads to higher privacy and utility preservation, and which may need future improvement.

4.4.1 Setup. For the ablation study, we focus on one mechanism (1-DIFFRACTOR) and two datasets (*SST2* and *Yelp*). For each dataset, we privatize all documents under the same setup as in the previous experiments, i.e., with the base epsilons of $\varepsilon \in \{0.1, 0.5, 1\}$, scaled to the average number of tokens per document. However, as opposed to before, we privatize each (dataset, ε) pair *five* times, each time with one method from our distribution toolkit disabled.

For the utility (*SST2*) and privacy (*Yelp*) ablation, we report the change in score (Δ), or how the corresponding score was affected by the disabling of the particular distribution scoring method. Thus, a more negative change in a metric, e.g., loss in utility, would imply that a given method is more effective when included than disabled.

The results of the ablation study are presented in Table 4.

5 Discussion

We critically reflect on the results presented in this work, as well as discuss opportunities and recommendations based on our findings.

5.1 More Privacy, Same Budget

An analysis of the experimental results begins with the strengths exhibited when performing DP text rewriting under our proposed distribution scheme rather than an equal, "naive" distribution.

As showcased in Table 1, using our toolkit leads to stronger protection against adversaries in attribute inference attacks (gender or authorship), in 15/18 static attacker scenarios and 13/18 adaptive attacker scenarios. These results are echoed in the Membership Inference evaluations, where our distribution outperforms naive distribution in nearly all of the *MTI* results, as well as all but one *NN* score. These results support the hypothesis that a more informed spending of the privacy budget in DP text rewriting can afford higher privacy levels given the same overall budget.

The implications of these results are clear. The provision of a certain privacy budget for a document leads to a particular privacy guarantee on paper (i.e., a DP guarantee), yet the *empirical* effects of such a guarantee can differ significantly depending on how the budget is "spent". These results show that in DP text rewriting, simply choosing an ε budget is not enough – a careful consideration of how this budget is allocated must also take place in order to maximize privacy protections in practice.

5.2 The Privacy-Utility Trade-off in Action

Naturally, a discussion of the privacy protections that our distribution method bolsters must also be discussed in light of its effect on the utility of the privatized data, or the *privacy-utility trade-off*.

Table 4: Ablation Study Results. Baseline scores represent the results using all five distribution methods, while */X* denotes the usage of four methods without *X*. Non-baseline values indicate the relative change (Δ, in %) from the baseline. Note that ablation results from the *masked token inference* (MTI) evaluation are not reported, due to non-significant changes. $^{*}std = 13.1$

ϵ	0.1						0.5						1.0					
	Baseline	/IC	/POS	/NER	/WI	/SD	Baseline	/IC	/POS	/NER	/WI	SD	Baseline	/IC	/POS	/NER	/WI	/SD
Utility (F1) ↑ (Δ)	$93.53_{0.0}$	+0.00	+0.00	+0.00	+0.00	+0.17	$94.01_{0.9}$	+0.21	-0.18	-0.48	-0.48	-0.48	$94.45_{0.6}$	-0.81	-0.92	-0.46	-0.92	-0.92
Adv. F1 (s) ↓ (Δ)	42.37	-1.56	-1.68	+0.34	-3.12	-3.55	55.09	-0.29	-0.52	+1.96	-0.99	-2.14	62.60	-0.87	+0.52	+1.56	-2.25	-2.08
Adv. F1 (a) ↓ (Δ)	$82.35_{0.9}$	+0.56	+1.73	+1.79	+3.97	+0.04	$88.44_{1.4}$	+1.68	+2.52	+3.41	-2.56	+0.42	$92.87_{0.7}$	-0.60	-0.94	-0.58	-1.16	-2.85
γ (s) ↑ (Δ)	1.81	+0.02	+0.03	-0.00	+0.05	-1.02	1.31	-0.12	+0.12	-0.02	+0.32	+0.34	0.91	+0.53	+0.59	+0.28	+0.63	+0.63
γ (a) ↑ (Δ)	1.21	-0.01	-0.03	-0.03	-0.06	-0.11	0.81	-0.17	+0.07	+0.04	+0.34	+0.30	0.46	+0.52	+0.60	+0.02	+0.60	+0.63
NN ↑ (Δ)	154	+4	-3	-27	+29	+32	17	+0	-2	-6	+6	+6	4	+0	+0	-1	+1	+3

(a) Yelp

ϵ	0.1						0.5						1.0					
	Baseline	/IC	/POS	/NER	/WI	/SD	Baseline	/IC	/POS	/NER	/WI	SD	Baseline	/IC	/POS	/NER	/WI	/SD
Utility (F1) ↑ (Δ)	$78.93_{0.2}$	+0.23	+0.79	+0.37	+0.40	-1.09	$85.31_{0.2}$	-9.31*	-1.03	+0.96	+0.43	-1.54	$88.94_{0.1}$	+0.05	+0.67	+0.82	+0.48	-3.28
CS ↑ (Δ)	0.459	+0.001	+0.010	+0.011	+0.001	-0.028	0.625	-0.001	+0.017	+0.019	-0.002	+0.058	0.721	+0.001	+0.022	+0.022	-0.001	+0.074
BLEU ↑ (Δ)	0.155	+0.000	+0.001	+0.003	+0.000	-0.005	0.203	+0.000	+0.007	+0.010	-0.002	-0.019	0.267	+0.001	+0.015	+0.023	-0.004	-0.40
PPL ↓ (Δ)	9849	+417	+514	+658	-2	+330	7266	+251	-259	-194	+470	+621	5813	+214	-142	+189	+77	+943

(b) SST2

In Table 3, a clear decrease in utility can be observed in nearly all cases of our versus naive distribution. On the surface, this utility loss is to be expected: if our aim is to privatize texts more rigorously by focusing on certain component tokens more than others, this will inevitably lead to a weaker semantic signal from the data. Interestingly, we observe that the effect on utility is different for our two chosen mechanisms. In this case of a word-level mechanism (1-Diffractor), the utility loss stays consistent and always entails a rather small loss. In the case of DP-BART, the effect on utility is clearly more severe, as demonstrated in the case of *BBC* and *IMDb*. This significant loss in utility is not absolute, though, as can be showcased in the *DocNLI* experiments, where our distribution with DP-BART performs the same or better than a naive distribution. These results highlight that budget distribution is not as clear-cut with document-level DP mechanisms, where segmenting inputs into sentences yields varying degrees of output quality.

In light of the infamous *privacy-utility trade-off*, we see that the relative consistency of the distributed 1-Diffractor utility loss is met with a generally lower capability to mitigate privacy risks. In particular, the results in Tables 1 and 2 show that data rewritten with 1-Diffractor, whether distributed with our method or not, is not as strong in protection against attribute or membership inference attacks. On the other hand, while DP-BART, particularly when distributed, can largely neutralize any privacy threat, the effect on utility is so significant that the trade-offs may be more similar to 1-Diffractor than meets the eye.

The trade-offs are most clearly demonstrated in Table 1 with the *Relative Gain* (γ) metric, which tells an interesting story. In defending against authorship attribution (*Yelp*), our distribution method nearly always (11/12) leads to more favorable trade-offs, and regardless of distribution method, DP-BART yields significantly higher relative gains. In contrast, the findings with gender identification (*Trustpilot*) are more mixed, with no clear winner regarding mechanism or distribution method. Beyond showing the complexity of the privacy-utility trade-off, these findings imply that considerations of budget distribution are also *task-specific*, and gains in terms of the trade-off do not come uniformly across different tasks.

5.3 When Does Distribution Make Sense? A Qualitative Analysis

Beyond the reported metrics in this work, one can look at side-by-side examples of DP-rewritten texts for insights.

Looking at selected examples in Table 5, one can begin to observe the differences in rewritten texts between the two distribution schemes. In the DP-BART example with $\varepsilon = 1500$, the naive distribution method not only fails to hide the "8 year old" cue, but it also magnifies this phrase in a later sentence. On the other hand, the rewritten text with our method makes no mention of this phrase, and in the case of $\varepsilon = 1000$, our rewritten text completely "masks" out the original sentence, albeit with a non-coherent replacement. Similarly, in the case of 1-Diffractor, we notice that more words are perturbed (changed) in the distributed examples rather than the naive. Moreover, certain writing cues, such as "We love it!", are never privatized in the naive distribution, but are finally considered in our distribution scheme (at $\varepsilon = 0.1$).

While these insights are anecdotal evidence, we hold that such differences are important in the consideration of text privatization. The examples illustrate the fact that in any given text, not all components of the greater whole are equally important, both semantically and from a privacy point of view; therefore, privacy budget allocation should follow the same logic. At the same time, the examples also demonstrate the pitfalls of a more informed distribution, such as in the non-coherent outputs of DP-BART at lower budgets. In addition, even with our proposed distribution, mechanisms such as 1-Diffractor struggle with truly obfuscating the original text, as in any case, significant semantic cues still remain. The qualitative analysis, therefore, teaches that while there is sense in distributing the privacy budget intelligently, there is still much work to be done. For more examples, we refer the reader to Table **??**.

5.4 Investigating the Distribution Methods

Following our ablation study (Table 4), we critically reflect on the merits and limitations of the individual distribution methods in our toolkit, leading to ideas and suggestions for further improvement.

Table 5: Selected Examples of Rewritten Texts from the *Yelp* Dataset, using the DP-BART Mechanism.

	ϵ		
		Original:	My 8 year old just LOVES it here - from musical instruments to jewelry to hand bags, everything is giftable and comes from a fair-trade community. It's great fun - but a bit pricey - and nothing here is a neccesity - but if you gotta buy gifts, at least here they are unique and helping another community. We love it!
DP-BART	1000	Naive	My 8 year old just loves it here - it's a great place to shop for gifts and toys - and the kids love it here! - from the start of the year - we have a lot of fun with it! - and it is a bit pricey - but if you gotta buy something here, it's worth it.My 8 yr old just LOVES it here
		Ours	Just a small-b%%% -%% e-the-w%%-%% The e-all-s%% It's also a great place to be if you want to be a part of the community - but it's also very difficult to be in the community.It's a great complete
	1500	Naive	My 8 year old loves this store - it's a gift shop that has everything you need for your kids. The kids are all over the place. The store is small - but the kids love it.My 8 yr old just LOVES this store. It's a great gift shop. The prices are great - but if you gotta buy a gift, this is a great
		Ours	This is the way we are going to go this year.We have a little bit of a way of doing this. The way we do it It's a bit pricey - but if you can't afford it, it's not a bad thing - and nothing here is cheap - but it's a little bit of a wonders

Methods that impact utility. We observe that, with slight deviations, all methods besides **SD** lead to an increase in utility when removed from the evaluations, thus suggesting that the utilization of these distribution methods leads to lower utility of the data. In the interesting case of **SD**, disabling this method actually leads to quite significant drops in performance (see the utility scores of *SST2* in the Table 4), which can plausibly be attributed to this method removing "outlier" tokens in the texts that may overtrain models to particular tokens. The utility-boosting properties of this method are also demonstrated in the other utility metrics, where, for example, disabling **SD** leads to the largest decreases in CS and BLEU.

Methods that impact privacy. The privacy results of the ablation study uncover an interesting dichotomy. In general, disabling any method besides **NER** seems to *improve* (lower) privacy scores in the *static* (s) attacker setting, whereas disabling **NER** always leads to *worse* (higher) results. This suggests that focusing the privacy budget on named entities is important in the static attacker setting.

In the adaptive attacker setting, other interesting findings arise. At lower privacy budgets (i.e., $\varepsilon = 0.1$ and $\varepsilon = 0.5$), *all* methods play an important role in reducing adversarial performance, except for one case (**WI** at $\varepsilon = 0.5$). However, at the higher budget setting ($\varepsilon = 1$), removing any given method only serves to improve (lower) the privacy results. This implies that budget distribution is most important in lower privacy budget regimes, where defending against more capable adversaries necessitates careful allocation of ε.

Similarly, the effect of certain methods is pronounced with lower privacy budgets, as showcased by the *NN* ablation scores. Here, we observe that **WI** and **SD** are influential against membership inference, whereas others do not play as large of a role.

As a final note regarding the low versus high privacy budgets, the relative gains (γ) of Table 4 illustrate that as the overall budget increases, it may make less sense (from a trade-off perspective) to distribute the budget with our method. This, however, would largely depend on whether balancing the trade-off is more important, as opposed to optimizing privacy (e.g., membership inference).

Main Takeaways. The results of the ablation study, in conjunction with the other results we present, show promise in the optimization of privacy budgets in DP text rewriting, while also highlighting important considerations going forward.

Our experiments present a cursory overview of the potential effectiveness of our proposed budget distribution methods, but the results merit further investigations. Taking **POS** as an example, we observe in Table 4 that this method generally contributes to better privacy scores, as showcased by the loss of privacy when it is disabled. However, this is met with increases in utility in some settings, and decreases in others. In this particular example, we cannot say with certainty whether the fixed weighting scheme of **POS** is optimal, and furthermore, exactly which weights can be adjusted. This discussion leads to the further consideration that while it is plausible that certain parts of speech are more relevant to privatization than others, we simply do not have the data to produce a more intelligent weighting scheme. This observation extends to our other proposed methods in the toolkit, where our initial assumptions about what is important in text privatization would be well-served to be backed by more informed data.

The various results presented in this work give credence to the complexity of privacy in textual data, as the many dimensions we present (i.e., the multiple angles of privacy and utility) make it difficult to definitively judge effectiveness in privatization. While this naturally calls for more work in privacy benchmarking and privacy metrics, it also sheds light on the subjective and individual nature of privacy in text. As an example, if privacy is strictly important in a certain data sharing scenario and one wishes to protect against strong adversaries, our budget distribution methods would be a very sensible choice. On the other hand, if utility is crucial while privacy is secondary, using a higher privacy budget without distribution might be the wiser choice. Although a continuation of this discussion is outside the scope of this work, the empirical results shown here certainly beckon for further such conversations.

6 Related Work

The field of DP text rewriting can be traced back to earlier works on authorship obfuscation using word-level Metric Differential Privacy [10]. Other works focusing on DP in NLP sought to improve word-level mechanisms, with later works tackling the challenges of utility preservation or efficiency [4, 6, 33, 45, 47, 50]. Later works transitioned to higher levels of syntactic hierarchy, such as with sentences [32] or document-level latent representations of text [2, 22, 23, 46]. DP text rewriting methods leveraging generative language models [13, 31, 42] have also been proposed in the recent literature as a way to produce more coherent privatized texts.

Researchers have also focused on identifying and addressing challenges in the field, especially at the core, where the integration of DP into the NLP realm is not immediately straightforward [16, 26]. Beyond clear challenges in the generation of coherent and utility-preserving privatized text [11, 31], questions of benchmarking and reproducibility [21, 36] have also been raised as important paths for future research. Finally, the meaning behind the guarantees that DP rewriting provides has also been a point of investigation [35, 43].

7 Conclusion

In this work, we investigate methods to improve the effectiveness of DP text rewriting by focusing on a more informed distribution of privacy budget amongst the tokens of a document. Given an input document and a fixed ε budget, we propose five methods and a scoring scheme to determine a sensible allocation of the budget to each of the document's components. In our conducted privacy experiments, we learn that in many cases, our proposed budget distribution leads to higher preserved privacy, against both attribute and membership inference attacks. At the same time, we observe that enhanced privacy does not come for free, as our budget distribution largely leads to lower utility in the privatized data.

Our findings highlight the importance of a more intelligent consideration of how a privacy budget is spent in DP text rewriting, resting upon the hypothesis that not all aspects of a text are equally as privacy-sensitive. We empirically demonstrate the privacy-utility trade-off at work, as well as qualitatively analyze the effects of budget distribution. Above all, our findings reveal that much work remains towards designing an optimal budget allocation scheme, and our proposed methods provide the groundwork for doing so.

As such, we propose that future work continues the discussion on the merits and challenges of informed privacy budget distribution in DP text rewriting. In particular, we hope that our proposed methods can be fine-tuned for better privacy protection, which would ideally be supported by user studies and a greater understanding of what it means to preserve privacy in textual data. Additionally, the extension of our work, both in distribution methods and rigorous testing on more DP mechanisms, would help to broaden the initial findings we present in this work.

References

[1] Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Guiding Text-to-Text Privatization by Syntax. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 151–162. https://doi.org/10.18653/v1/2023.trustnlp-1.14

[2] 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*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 3997–4007. https://doi.org/10.18653/v1/2021.naacl-main.314

[3] 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)*. USENIX Association, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[4] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 883–890. https://doi.org/10.1137/1.9781611977653.ch99

[5] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13*. Springer, 82–102. https://doi.org/10.1007/978-3-642-39077-7_5

[6] Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5747–5758. https://doi.org/10.18653/v1/2023.findings-acl.355

[7] Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5747–5758. https://doi.org/10.18653/v1/2023.findings-acl.355

[8] 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12. https://doi.org/10.1007/11787006

[10] Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*. Springer International Publishing, 123–148. https://doi.org/10.1007/978-3-030-17138-4_6

[11] Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021. Research Challenges in Designing Differentially Private Text Generation Mechanisms. In *The International FLAIRS Conference Proceedings*, Vol. 34. https://doi.org/10.32473/flairs.v34i1.128461

[12] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 178–186. https://doi.org/10.1145/3336191.3371856

[13] James Flemings and Murali Annavaram. 2024. Differentially Private Knowledge Distillation via Synthetic Text Generation. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 12957–12968. https://doi.org/10.18653/v1/2024.findings-acl.769

[14] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 42–47. https://aclanthology.org/P11-2008

[15] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access* 11 (2023), 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

[16] 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*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1522–1528. https://doi.org/10.18653/v1/2021.emnlp-main.114

[17] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[18] Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User Review Sites as a Resource for Large-Scale Sociolinguistic Studies. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 452–461. https://doi.org/10.1145/2736277.2741141

[19] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* 54, 11s, Article 235 (Sept. 2022), 37 pages. https://doi.org/10.1145/3523273

[20] Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially Private Natural Language Models: Recent Advances and Future Directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 478–499. https://aclanthology.org/2024.findings-eacl.33

[21] Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-Rewrite: Towards Reproducibility and Transparency in Differentially Private Text Rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao

Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2927–2933. https://aclanthology.org/2022.coling-1.258

[22] Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for Privatized Text Rewriting under Local Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13914–13934. https://doi.org/10.18653/v1/2023.findings-acl.874

[23] Timour Igamberdiev, Doan Nam Long Vu, Felix Kuennecke, Zhuo Yu, Jannik Holmer, and Ivan Habernal. 2024. DP-NMT: Scalable Differentially Private Machine Translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 94–105. https://aclanthology.org/2024.eacl-demo.11

[24] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2008. What Can We Learn Privately?. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. 531–540. https://doi.org/10.1109/FOCS.2008.27

[25] T.P. Klammer, M.R. Schulz, and A.D. Volpe. 2010. *Analyzing English Grammar*. Longman. https://books.google.de/books?id=1dbRPgAACAAJ

[26] Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential Privacy in Natural Language Processing The Story So Far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, Oluwaseyi Feyisetan, Sepideh Ghanavati, Patricia Thaine, Ivan Habernal, and Fatemehsadat Mireshghallah (Eds.). Association for Computational Linguistics, Seattle, United States, 1–11. https://doi.org/10.18653/v1/2022.privatenlp-1.1

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[28] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. arXiv:2308.03281 [cs.CL] https://arxiv.org/abs/2308.03281

[29] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4188–4203. https://doi.org/10.18653/v1/2021.acl-long.323

[30] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 142–150. https://aclanthology.org/P11-1015

[31] Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The Limits of Word Level Differential Privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 867–881. https://doi.org/10.18653/v1/2022.findings-naacl.65

[32] Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. Sentence-level Privacy for Document Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3367–3380. https://doi.org/10.18653/v1/2022.acl-long.238

[33] Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024. 1-Diffractor: Efficient and Utility-Preserving Text Obfuscation Leveraging Word-Level Metric Differential Privacy. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics* (Porto, Portugal) *(IWSPA '24)*. Association for Computing Machinery, New York, NY, USA, 23–33. https://doi.org/10.1145/3643651.3659896

[34] Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024. A Collocation-based Method for Addressing Challenges in Word-level Metric Differential Privacy. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, Ivan Habernal, Sepideh Ghanavati, Abhilasha Ravichander, Vijayanta Jain, Patricia Thaine, Timour Igamberdiev, Niloofar Mireshghallah, and Oluwaseyi Feyisetan (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 39–51. https://aclanthology.org/2024.privatenlp-1.5

[35] Stephen Meisenbacher and Florian Matthes. 2024. Thinking Outside of the Differential Privacy Box: A Case Study in Text Privatization with Language Model Prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5656–5665. https://doi.org/10.18653/v1/2024.emnlp-main.324

[36] Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024. A Comparative Analysis of Word-Level Metric Differential Privacy: Benchmarking the Privacy-Utility Trade-off. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 174–185. https://aclanthology.org/2024.lrec-main.16

[37] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. https://doi.org/10.48550/arXiv.2311.17035 arXiv:2311.17035 [cs.LG]

[38] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy Risks of General-Purpose Language Models. In *2020 IEEE Symposium on Security and Privacy (SP)*. 1314–1331. https://doi.org/10.1109/SP40000.2020.00095

[39] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[40] Jonathan Schler, Moshe Koppel, and Shlomo Argamon. 2006. Effects of age and gender on blogging.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199–205. https://aaai.org/papers/0039-ss06-03-039-effects-of-age-and-gender-on-blogging/

[41] Samuel Sousa and Roman Kern. 2023. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review* 56, 2 (2023), 1427–1492. https://doi.org/10.1007/s10462-022-10204-6

[42] Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8442–8457. https://doi.org/10.18653/v1/2023.findings-emnlp.566

[43] Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. 2024. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 507–527. https://doi.org/10.18653/v1/2024.findings-emnlp.29

[44] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (Eds.). Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[45] 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*. 305–314. https://doi.org/10.1145/3209978.3210008

[46] 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* (, Virtual Event, Lyon, France,) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 721–731. https://doi.org/10.1145/3485447.3512232

[47] Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021. Density-aware differentially private textual perturbations using truncated Gumbel noise. In *The International FLAIRS Conference Proceedings*, Vol. 34. https://doi.org/10.32473/flairs.v34i1.128463

[48] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. https://doi.org/10.48550/arXiv.2403.05156 arXiv:2403.05156 [cs.CR]

[49] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4913–4922. https://doi.org/10.18653/v1/2021.findings-acl.435

[50] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 3853–3866. https://doi.org/10.18653/v1/2021.findings-acl.337