

Just Rewrite It Again: A Post-Processing Method for Enhanced Semantic Similarity and Privacy Preservation of Differentially Private Rewritten Text

Stephen Meisenbacher

Technical University of Munich

School of Computation, Information and Technology

Department of Computer Science

Garching, Germany

stephen.meisenbacher@tum.de

Florian Matthes

Technical University of Munich

School of Computation, Information and Technology

Department of Computer Science

Garching, Germany

matthes@tum.de

ABSTRACT

The study of Differential Privacy (DP) in Natural Language Processing often views the task of text privatization as a *rewriting* task, in which sensitive input texts are rewritten to hide explicit or implicit private information. In order to evaluate the privacy-preserving capabilities of a DP text rewriting mechanism, *empirical privacy* tests are frequently employed. In these tests, an adversary is modeled, who aims to infer sensitive information (e.g., gender) about the author behind a (privatized) text. Looking to improve the empirical protections provided by DP rewriting methods, we propose a simple post-processing method based on the goal of aligning rewritten texts with their original counterparts, where DP rewritten texts are rewritten *again*. Our results show that such an approach not only produces outputs that are more semantically reminiscent of the original inputs, but also texts which score on average better in empirical privacy evaluations. Therefore, our approach raises the bar for DP rewriting methods in their empirical privacy evaluations, providing an extra layer of protection against malicious adversaries.

CCS CONCEPTS

- Security and privacy → Domain-specific security and privacy architectures;
- Computing methodologies → Natural language processing.

KEYWORDS

Data Privacy, Differential Privacy, Natural Language Processing

ACM Reference Format:

Stephen Meisenbacher and Florian Matthes. 2024. Just Rewrite It Again: A Post-Processing Method for Enhanced Semantic Similarity and Privacy Preservation of Differentially Private Rewritten Text. In *The 19th International Conference on Availability, Reliability and Security (ARES 2024), July 30-August 2, 2024, Vienna, Austria*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664476.3669926>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ARES 2024, July 30-August 2, 2024, Vienna, Austria

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1718-5/24/07

<https://doi.org/10.1145/3664476.3669926>

1 INTRODUCTION

The proliferation of Large Language Models (LLMs) in recent years has given rise to discussions of data privacy in Natural Language Processing (NLP), particularly as the need for high-quality user-generated text becomes increasingly important to fuel the training of such models [3]. While LLMs have demonstrated very impressive capabilities across the spectrum of NLP tasks, the privacy risks inherent in the requirement for massive amounts of training data have motivated the study of privacy-preserving NLP [20, 30]. This need is made even more salient when considering the sheer amounts of data being passed to hosted LLMs as prompts [9].

In response to these concerns, a stream of research within the NLP community studies the integration of Differential Privacy (DP) [8] into NLP workflows. As a mathematically grounded blueprint for achieving privacy in data processing scenarios, DP offers a promising solution, yet its direct incorporation into the textual domain does not come without challenges [11, 20, 24]. Nevertheless, many innovative solutions have been proposed in recent works [17], among these the idea of *differentially private text rewriting*.

In a DP text rewriting scenario, an input text is transformed under DP guarantees with the help of a *rewriting mechanism*, which ideally outputs a privatized text that is semantically similar, yet obfuscated from the original [24]. These mechanisms often operate at the *local* level, where users rewrite their data before releasing it to some aggregator. Different rewriting mechanisms operate at various syntactic levels, such as at the word level [4, 12], sentence level [26], or document level [19]. In any scenario, the level at which a DP rewriting mechanism operates leads to the DP guarantee that is provided: for example, given a DP privatized *sentence* in the local setting, this sentence is indistinguishable from all other sentences within some bound, governed by the DP privacy parameter ϵ .

An important part of designing DP rewriting mechanisms is the evaluation of its privacy-preserving capabilities. In many recent works, this evaluation takes the form of *empirical privacy* tests, where the DP privatized texts are shown to reduce the ability of an attacker to perform some adversarial inference task, as compared to the non-privatized (unwritten) baseline. In modeling such adversaries, two basic archetypes have been predominantly used in the literature, namely the *static* and *adaptive* attackers [24, 34, 37].

As shown by empirical privacy evaluations in the literature, it is often the case that the *adaptive* attacker proves to be a considerably more difficult challenge for DP rewriting mechanisms [34], as this attacker is able to mimic the rewriting process and thereby train

a more accurate adversarial model. At the same time, achieving better results in empirical privacy evaluations often necessitates lower (stricter) ϵ values, which in turn can lead to loss of utility, i.e., semantic similarity to the original texts. This highlights a major challenge of DP text rewriting, that is, finding the balance between privacy and utility [19, 28].

In this work, we aim to address the challenge presented by the strong capabilities of adversaries modeled in the literature, particularly with the adaptive attacker. To accomplish this goal, we aim to leverage an important property of DP, the *post-processing* principle, to propose a method in which users can enhance the privacy preservation of their DP rewritten texts. As such, we pose the following research question:

How can users in the local differentially private text rewriting scenario leverage language models to enhance both the empirical privacy and semantic similarity of their rewritten texts?

To answer this question, we design, formulate, and evaluate a post-processing method that essentially rewrites the DP rewritten text *again*, with the goal of increasing its privacy while also aligning its semantics to the original text. In evaluating our proposed method, we observe that this extra post-processing step provides clear and significant privacy gains, while also often resulting in higher semantic similarity to the original text counterparts.

We make three contributions to the field of DP text rewriting:

- (1) To the best of the authors' knowledge, this is the first work to leverage the post-processing property of DP to improve the privacy and quality of DP rewritten texts.
- (2) We present a mechanism-agnostic method which demonstrates strong capabilities to enhance the privacy preservation and semantic similarity of DP rewritten texts.
- (3) We add to the body of knowledge on DP text rewriting evaluation by highlighting the usefulness of a post-processing step in enhancing the abilities of existing DP mechanisms.

2 FOUNDATIONS

In order to ground our work in previous literature, we now walk through key foundational concepts, which become important in motivating our proposed method.

2.1 Differentially Private Text Rewriting

The goal of differentially private text rewriting is to rewrite a sensitive input text in a manner that satisfies Differential Privacy (DP) [8]. Specifically, a mechanism \mathcal{M} with privacy parameter ϵ satisfies DP if for any two *adjacent* inputs x and y , and $\epsilon > 0$:

$$\frac{\Pr[\mathcal{M}(x) = z]}{\Pr[\mathcal{M}(y) = z]} \leq e^\epsilon \quad (1)$$

In essence, the inequality in Equation 1 necessitates a certain level of *indistinguishability* between the outputs of two neighboring inputs. The notion of *adjacent* or *neighboring* is dataset-specific, but must be defined in order to satisfy the original notion of DP as defined in Equation 1. This level is governed by the ϵ parameter: a higher ϵ requires less indistinguishability, and vice versa.

The primary challenge with DP text rewriting comes with the design of the underlying *mechanism* that performs the rewriting [17,

20]. In light of the considerations required by DP, important design decisions include the definition of what *any two text inputs* means, often referred to as *adjacency*. In the literature, adjacency is often either defined on the token-/word-level or the sentence-/document-level. Both approaches have advantages and drawbacks: word-level approaches suffer from lack of contextualization [24], while offering the ability for tighter privacy guarantees. Mechanisms operating on entire documents can better preserve syntactic and semantic coherence, but document representations in large dimensions often necessitate high levels of noise to achieve the DP guarantee, due to their large *sensitivity* [19].

In this work, we focus on both approaches, namely ones that provide either word- or document-level DP guarantees. We are thereby able to evaluate the effectiveness of our proposed evaluation on both streams of DP text rewriting research. The exact mechanisms we utilize will be introduced in Section 4, but firstly, two important notions associated with DP text rewriting are introduced.

Local Differential Privacy. DP rewriting mechanisms often leverage the *local* notion of Differential Privacy. As opposed to *global* or *central* DP, Local DP (LDP) places the utilization of DP mechanisms at the user level [7]. In other words, the transformation (rewriting) of text data is performed by the user locally before releasing the output to some central aggregator. The definition of ϵ -LDP is as follows, for some mechanism π , *any* inputs x and x' , and $\epsilon > 0$,

$$\frac{\Pr[\pi(x) = z]}{\Pr[\pi(x') = z]} \leq e^\epsilon \quad (2)$$

As one may observe, the important difference that comes with LDP is the requirement for *any two* inputs from the user to satisfy Equation 2. Although LDP allows for privacy protection to be ensured already at the user level, the major drawback comes with this strict adjacency requirement, thereby often leading to higher amounts of additive noise needed, including in the text rewriting scenario [19]. In this work, we seek to improve this trade-off, whereby our proposed method offers better utility than the raw rewritten outputs of DP text rewriting mechanisms.

Post-processing. An important property of Differential Privacy that is leveraged in this work is the notion of *post-processing*. In particular, mechanisms that satisfy DP (i.e., satisfy Equation 1) are robust against post-processing, defined as any arbitrary operation or computation performed on top of the output of a DP mechanism is safe. Specifically, given a DP mechanism \mathcal{M} , some deterministic or randomized function g , and input x , $g(\mathcal{M}(x))$ upholds the guarantee provided by \mathcal{M} [8].

This important property states that the output of any function g is still resistant against adversaries, despite these adversaries possessing some auxiliary information. From an information theory perspective, this is due to the fact that regardless of what *auxiliary* knowledge may be possessed by an adversary, lacking knowledge of the private database (here, texts) makes it impossible to compute the function of the outputs of \mathcal{M} [8]. As stated by Near and Abuah [29], DP mechanisms often leverage this useful DP property to improve the accuracy of DP outputs. As such, we aim to perform post-processing on DP rewritten texts with the goal of enhancing the utility and usability of these texts.

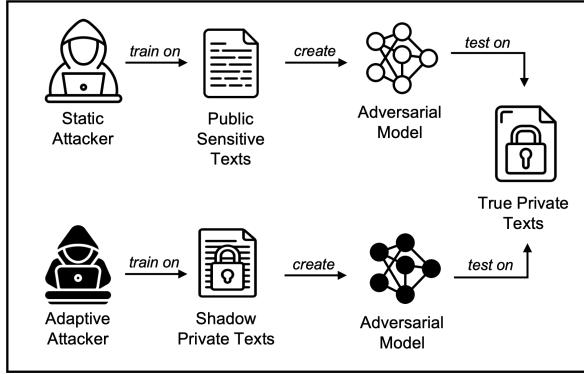


Figure 1: The *Static* and *Adaptive* Attackers. The adaptive attacker, with knowledge of the rewriting mechanism, generates “shadow” texts by rewriting publicly available texts from the same domain as the target “true” private texts.

2.2 Empirical Privacy Evaluation

The evaluation of DP text rewriting mechanisms offer takes the form of two-part experiments, resting the *utility* and *privacy* preservation of a proposed mechanism. In particular, a good rewriting mechanism not only protects the privacy of a text, but only rewrites the original text in a way that preserves its downstream utility, i.e., its ability to be useful for training models in some defined task [24].

With this, the question becomes how to evaluate the “privacy preservation” of a rewriting mechanism. In many recent works, such an evaluation takes the form of *empirical privacy* experiments, where the demonstration of a mechanism’s privacy-preserving capabilities is performed empirically [17, 24, 34]. At a high level, this is typically done by showing that the output texts (i.e., post-rewriting) reduce the adversarial advantage of an attacker seeking to misuse the text for some nefarious purpose. Prominent examples, which we employ later in this work, include inferring the gender or authorship of the writer behind a given text. As such, a good DP rewriting mechanism should produce texts that reduce an attacker’s ability to perform such inferences accurately.

To perform these empirical privacy experiments, two types of attackers have been modeled by the recent literature: the *static* and *adaptive* attackers [24]. Both types of attackers work towards the ultimate goal of accurately performing inferences (for a sensitive attribute such as gender) given some corpus of DP-rewritten texts:

- **Static attacker:** the static attacker has access to the DP rewritten texts, but has no knowledge of the mechanism used to perform the rewriting. The attacker does, however, possess knowledge of the *domain* of the original data, and furthermore, has access to a public dataset of (unrewritten) texts associated with the target sensitive attribute. Using this public data, the static attacker trains a model to predict the target attribute given an input text, and uses this model to infer the attribute of each text in the DP rewritten corpus.
- **Adaptive attacker:** the adaptive attacker possesses all the knowledge that the static attacker does. In addition, the adaptive attacker is stronger in the sense that the exact mechanism (including ϵ) is known. Using this knowledge, which

includes the ability to run the mechanism, the adaptive attacker uses the public data to create a DP rewritten version of this data. The attacker then trains a model on the *rewritten* data to predict the target attribute, i.e., to predict the sensitive attribute of the original DP rewritten texts.

Both the static and adaptive attacker setups are illustrated in Figure 1. As shown unanimously by recent works [17, 24, 34], the adaptive attacker proves to be the stronger adversary. This can be explained by the fact that the distribution on which the adversarial model is trained more closely matches that of the target texts.

In this work, we use both of these attacker types to underline our empirical privacy evaluations. Specifically, we introduce a post-processing method which aims to decrease the adversarial advantage of both attackers, in the way that our post-processing method provides an extra layer of protection on top of DP rewritten texts. This method is introduced in Section 3.

2.3 Text2Text Generation

Recent advances in language models have demonstrated the impressive abilities of these models to generate highly coherent and plausible texts given an input prompt [42]. Many (large) language models also show very strong performance when *fine-tuned* for some specific downstream task, which necessitates only further training data upon which a base model can be improved.

A prominent paradigm of language model fine-tuning comes in the form of *Text2Text Generation* [23], which describes the case where there is a full-text input and full-text output. Common tasks employing such as setup include Machine Translation, Text Summarization, Text Simplification, and Paraphrasing, among others. Given *parallel* datasets, modern LMs can be fine-tuned efficiently, where the base model is then made an “expert” by being given domain- and/or task-specific training data.

In this work, we leverage the Text2Text fine-tuning setup to envision a post-processing step to “rewrite” texts “again”, that is effectively to improve the empirical privacy gains of DP text rewriting by once again rewriting these texts. The definition and requirements of such a method are now introduced in the following.

3 A POST-PROCESSING METHOD FOR DP REWRITTEN TEXTS

In the following, we outline in detail our proposed post-processing method, which can be broken down into two tracks. We present these methods as a way to improve both the semantic similarity and privacy preservation offered by DP rewritten texts. The process flow of these two tracks is illustrated in Figure 2.

3.1 Preliminaries

We assume a user wishing to use some DP rewriting mechanism in the LDP setting. Concretely, a user will rewrite his or her textual dataset before releasing it to some data aggregator or central analyst. To do this, the user leverages a DP mechanism M with privacy budget ϵ , where this budget is chosen for the entire text dataset to be rewritten. Additionally, we assume that the user has access to large-scale public text corpora, as well as the ability and resources to fine-tune LLMs on such data in a Text2Text Generation setup.

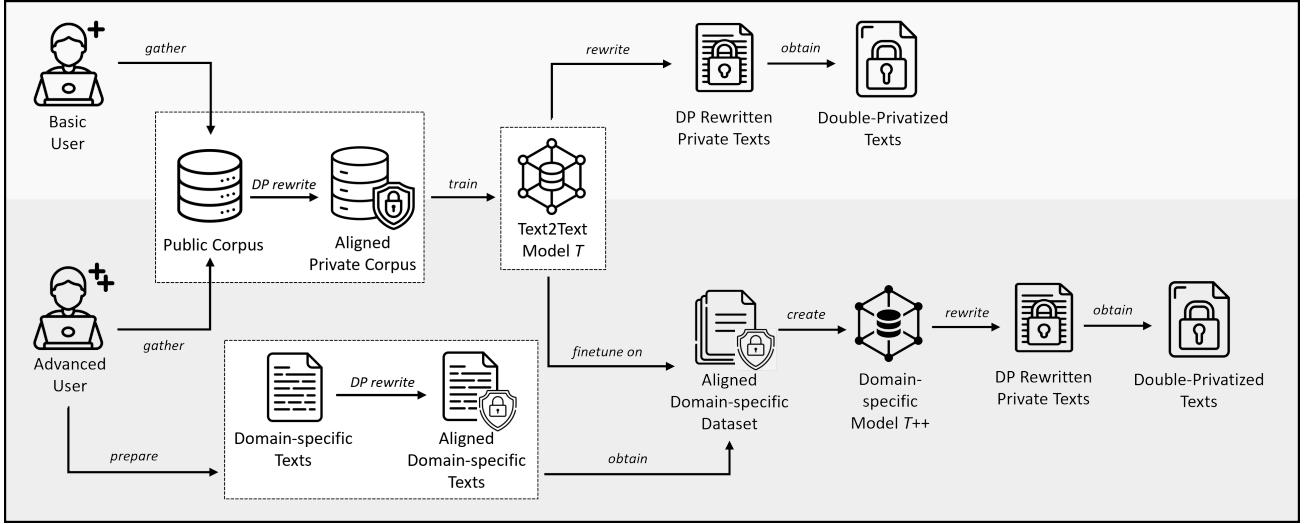


Figure 2: A post-processing pipeline to improve the semantic coherence and privacy preservation of DP rewritten text. Both “tracks” involve a user creating a model T to rewrite DP rewritten text again. While a basic user utilizes large-scale public text corpora, the more advanced user (bottom track) also leverages domain-specific data to fine-tune T further to obtain T^{++} .

The motivation of the user is to re-align DP rewritten output text to be semantically closer to the original meaning of the text, thereby boosting the utility and fidelity of the text eventually to be released. As it turns out, an added benefit of performing this extra step also is the enhanced privatization of the output text, providing an added incentive for the user. These benefits will be empirically demonstrated in Section 5 and analyzed in Section 6.

3.2 Method

The following outlines the basic steps taken by the user in our proposed post-processing pipeline to rewrite the DP rewritten texts once again. The ensuing narrative then goes into detail on each of these steps. The user takes the following steps:

- (1) Collect a large number of texts from a public text corpus
- (2) Use the rewriting mechanism \mathcal{M} with parameter ϵ to rewrite the public corpus into a *aligned private corpus*
- (3) Finetune a Text2Text LM to generate text in the *opposite* direction, i.e., generate the public corpus text given the aligned corpus text as input
- (4) Use the Text2Text LM to “re-rewrite” the DP rewritten texts
- (5) Release the doubly rewritten texts

In a more advanced setup, the user takes the same steps (1)-(3), but then continues as follows:

- (4) Generate *aligned private texts* from the same domain as the DP rewritten texts
- (5) Further finetune the Text2Text LM on this parallel domain-specific data, once again in the “reverse” direction
- (6) Use the further finetuned Text2Text LM to “re-rewrite” the DP rewritten texts
- (7) Release the doubly rewritten texts

3.2.1 Corpus Creation. The first step of our proposed post-processing pipeline comes with the collection of a large collection of text samples from a publicly available corpus, such as Wikipedia [13] or Common Crawl [31]. Using these text samples, a user can use the DP rewriting mechanism \mathcal{M} to rewrite the samples in DP rewritten versions, thus creating a parallel, *aligned* corpus.

For the purposes of this work, we utilize a random sample of 100,000 English text samples from C4 (Colossal Cleaned Crawled Corpus), made available by Raffel et al. [31] and cleaned by allenai.¹.

3.2.2 LM Fine-Tuning. Given the prepared aligned corpus, the next task is to fine-tune a Text2Text LM. Much like in the tasks of Machine Translation or Text Simplification, such a fine-tuning task requires a set of parallel *source-target* text pairs, which are naturally provided in the DP rewriting process. With this, the DP rewritten texts become the *source* documents, and the original texts represent the *target* documents. This setup is then used to fine-tune an LM, with the effective aim of “shifting” the rewritten texts back to the general semantic makeup of the original texts.

3.2.3 Track Two: Further Fine-Tuning. As introduced in the above steps, a more advanced user may opt to fine-tune the Text2Text model even further, given the presence of domain-specific texts that more closely mirror the true texts to be released. The motivation behind this is that large-scale public text corpora may contain a vast variety of text subjects, which may not be completely suitable given a set of sensitive texts in one specific domain, that must, in turn, be privatized (rewritten).

With this domain-specific data, the user would proceed as before with the public corpus, i.e., first generate the parallel, aligned dataset, and then proceed to fine-tune the previously obtained Text2Text model further. The output is thus a model that is more

¹<https://huggingface.co/datasets/allenai/c4>

tuned to the target domain, with the ability to recapture the semantics in a particular subject area.

A representative example of domain-specific data, one that we leverage and evaluate later in this work, is that of *user reviews*. If a user is to privatize his or her personal reviews before releasing them, publicly available datasets such as Yelp Reviews [41] or Trustpilot Reviews [16] may be useful.

3.2.4 Data Rewriting and Release. The final step, given the (domain-specific) fine-tuned LM, is to use the model to rewrite the DP rewritten text outputs. Only these “doubly” rewritten text outputs are then shared with third parties, and both the original and DP rewritten texts remain private to the user.

4 EXPERIMENTAL SETUP

We describe our experimental setup, which includes the selected DP rewriting mechanisms, design choices for our proposed method, and empirical privacy experiment parameters.

4.1 Selected DP Mechanisms

For the evaluation of our proposed method, we choose two DP rewriting mechanisms from the recent literature.

DP-BART. The DP-BART mechanism [19] was proposed as an LDP mechanism using the BART model architecture [22] as its basis. In essence, DP calibrated noise is added to the latent text representation existing between the encoder and decoder components of BART. As such, the rewriting mechanism operates at the *document-level*, where a single output document is generated in a DP manner given an input document. In particular, we utilize the **DP-BART-CLV** variant as proposed by Igamberdiev and Habernal [19], which clips the latent vector values before adding noise.

For ϵ values, we choose the first and third quartile values from the range of values evaluated by the original authors. This corresponds to the ϵ values of 625 and 1875.

DP-Prompt. Leveraging the proxy task of *paraphrasing*, DP-PROMPT [34] presents a method to generate privatized documents under LDP guarantees by using (large) language models as paraphrasers. Utilizing a temperature sampling mechanism as a form of Exponential Mechanism [25], privatized documents are generated word-by-word by a DP sampling of the next token, thus providing DP guarantees at the *word-level*. For our implementation of DP-PROMPT, we use the FLAN-T5-LARGE language model (780M parameters) [6].

As with DP-BART, we use two ϵ values as used by the authors of DP-PROMPT, namely 137 and 206. Specifically, following Utpala et al. [34], we first measure the logit range of our selected FLAN-T5-LARGE model. This is done by estimating the range via running the model on 100 randomly selected texts from the C4 corpus, and accordingly choosing the clipping range to be $(\text{logit}_{\min}, \text{logit}_{\max}) = (-95, 8)$, thus leading to a sensitivity of 103. Next, we choose the temperature values T of 1.0 and 1.5 (first and third quartile values), which correspond to the two ϵ values listed above².

4.2 Post-Processing Pipeline

Datasets. As stated in Section 3.2.1, we use a random sample of 100,000 text samples from the C4 Corpus to serve as the *public corpus* as envisioned in our post-processing pipeline. Additionally, we use two *domain-specific* datasets, namely the Yelp and Trustpilot reviews, which will be covered in more detail in the following outline of the empirical privacy experimental setup.

Model Fine-Tuning. In both fine-tuning scenarios, that is the base fine-tuning to create model T and the further fine-tuning to obtain $T++$, we once again employ the FLAN-T5-LARGE model checkpoint. Given the input parallel corpus (public-private or domain-specific), the model is trained for one epoch with a learning rate of 5e-5, and otherwise all default HuggingFace Trainer³ parameters. This process resulted in a total of 12 trained models, namely:

Basic User:

- (1) Model T , fine-tuned on *aligned private corpus (apc)* with DP-BART ($\epsilon=625$)
- (2) Model T , fine-tuned on *apc* with DP-BART ($\epsilon=1875$)
- (3) Model T , fine-tuned on *apc* with DP-PROMPT ($\epsilon=137$)
- (4) Model T , fine-tuned on *apc* with DP-PROMPT ($\epsilon=206$)

Advanced User:

- (1) Model $T++$, further fine-tuned on the Yelp *aligned domain-specific corpus* with DP-BART ($\epsilon=625$)
- (2) Model $T++$, further fine-tuned on the Yelp *aligned domain-specific corpus* with DP-BART ($\epsilon=1875$)
- (3) Model $T++$, further fine-tuned on the Yelp *aligned domain-specific corpus* with DP-PROMPT ($\epsilon=137$)
- (4) Model $T++$, further fine-tuned on the Yelp *aligned domain-specific corpus* with DP-PROMPT ($\epsilon=206$)
- (5) Model $T++$, further fine-tuned on the Trustpilot *aligned domain-specific corpus* with DP-BART ($\epsilon=625$)
- (6) Model $T++$, further fine-tuned on the Trustpilot *aligned domain-specific corpus* with DP-BART ($\epsilon=1875$)
- (7) Model $T++$, further fine-tuned on the Trustpilot *aligned domain-specific corpus* with DP-PROMPT ($\epsilon=137$)
- (8) Model $T++$, further fine-tuned on the Trustpilot *aligned domain-specific corpus* with DP-PROMPT ($\epsilon=206$)

4.3 Empirical Privacy Experiments

With the post-processing pipeline, the goal is now to evaluate its effect on the *empirical privacy* protection provided by DP rewriting mechanisms. Additionally, we also measure the effects of the post-processing step on the *utility* of the text, an important counterbalance to be measured in DP text rewriting [24, 28].

4.3.1 Datasets. As mentioned in Section 4.2, we employ two datasets, both of which allow for direct empirical privacy measurement.

Yelp Reviews. The Yelp Review dataset [41] is a dataset containing user reviews from the popular Yelp platform. Each review contains a star rating from 1-5, leading to the dataset often being used for sentiment analysis. In particular, we use the same data subset as used by Utpala et al. [34], which contains 17,295 reviews from 10 frequent users. Thus, we model the empirical privacy task as an

²Rounded values from the formula $\epsilon = \frac{2\Delta}{T}$, where Δ represents the sensitivity.

³https://huggingface.co/docs/transformers/en/main_classes/trainer

authorship identification task, wherein an adversary attempts to infer the author given a written text.

Trustpilot Reviews. The Trustpilot Review dataset [16] is a large corpus of reviews from the Trustpilot platform. As with Yelp, each Trustpilot review is scored from 1-5. Additionally, the dataset also lists the gender of each reviewer, thus leading to the *gender identification* adversarial task. As the original dataset is quite expansive, we take a random 10% of reviews with the gender listed, resulting in an evaluation dataset of 29,490 reviews.

4.3.2 Modeling Adversaries. As introduced in Section 2.2, we model two types of adversaries for the empirical privacy evaluation: the *static* and *adaptive* attackers. In order to mimic the data available that these attackers would leverage we perform the following steps:

- (1) For each dataset (Yelp/Trustpilot), we take a 90-10 train-validation split, using a random seed of 42.
- (2) The train split is known and used by the attacker. The static attacker only has access to the original version of the texts, while the adaptive attacker has access to the privatized (rewritten) versions.
- (3) The validation split represents the “true” private text, i.e., the text which the user is releasing, and which the user aims to protect further using post-processing. This text, released in privatized form, is the target of both adversaries.

Given the train split as described above, the attacker in question trains an adversarial model to infer the sensitive attribute, i.e., author or gender. The static attacker trains on the original (clean) texts, while the adaptive attacker trains on the privatized texts (rewritten by mechanism \mathcal{M}).

To build these adversarial classifiers, a DEBERTA-v3-BASE model [14, 15] is fine-tuned. The model is trained with a 10-class classification head for the authorship identification task, and a 2-class head for gender identification. Training is run for one epoch on a given dataset with a learning rate of $5e-5$ and otherwise default training parameters. With the static attacker, a single model is trained and then evaluated on all DP rewritten variants, while for the adaptive attacker, a model is trained for each variant and subsequently evaluated on the corresponding privatized validation set. All model training is performed on a single NVIDIA RTX6000 48GB GPU.

4.4 Metrics

Empirical privacy is measured by the ability of a mechanism’s rewritten text to *reduce* the adversarial advantage of a given attacker. In this study, this is represented by the reduction in F1 score of the attacker evaluating an adversarial model on the private validation split. This reduction is presented against the baseline of training and testing on the clean, non-privatized version of the data, i.e., what an attacker could achieve given full access to the entire target dataset. As such, a lower F1 score represents a higher empirical privacy result. It is important to note that all scores represent an average of three runs, that is, a model is trained three times on different shuffles of the dataset, and the three evaluation results are averaged for the final score.

We also measure the preserved *utility* of the (doubly) rewritten text data, modeled as the *semantic similarity* between the (*original*,

rewritten) and (*original, doubly rewritten*) pairs of text. To measure semantic similarity, we utilize Sentence Transformer models [32], specifically the ALL-MPNET-BASE-V2 and ALL-MINILM-L6-V2 encoder models. The cosine similarity between the encoded pairs of texts is taken and subsequently averaged over an entire rewritten dataset. Since two models are used to account for model-specific differences, the cosine similarity (CS) scores reported represent the average score between the two models.

5 EMPIRICAL RESULTS

The results of our empirical privacy experiments, namely on the Yelp and Trustpilot datasets, are given in Table 1.

As noted in Table 1, we present the empirical privacy results in the form of *adversarial advantage*, or specifically the F1 score achieved by an adversary (static or adaptive) in all selected scenarios. The baseline scenario depicts the performance an attacker could achieve given full access to the original, non-privatized data. Then, given a (*mechanism, ϵ*) pair, we present the empirical privacy results for each rewriting scenario: (1) base output from the DP rewriting mechanism, (2) double-privatized outputs from our “basic user”, and (3) double-privatized outputs from our “advanced user”.

The results also include the CS score, as introduced previously, which captures the degree to which semantic meaning is kept from the original text to the rewritten text. Thus, we pose that a higher CS score indicates a closer preservation of original semantic meaning.

6 DISCUSSION

We now analyze in-depth the empirical results presented in Table 1, as well as discuss the merits and limitations of our method.

6.1 The Benefits of Post-Processing

An analysis of the empirical results reveals the strengths of our post-processing method, particularly in reducing adversarial performance. Concretely, in the 16 attacker scenarios presented (8x static, 8x adaptive), either our *basic* or *advanced* method leads to the lowest adversarial performance in 13 of the 16 scenarios. In some cases, particularly with DP-BART, the adaptive attacker’s performance is nearly 50% lower as compared to the DP rewritten texts without post-processing. Similarly, our method also proves to be quite useful in reducing the performance of the static attacker.

Looking to both the static and adaptive scores, one can note that only the texts resulting from our methods can truly *neutralize* the adversarial advantage. In other words, these attackers achieved scores equal to or worse than *majority-class guessing*, i.e., simply choosing one class known to be the most frequent. In the Yelp dataset, the most frequent author writes 17.5% of the reviews, and in Trustpilot, the split is 57.9%/42.1% for male/female. As can be seen in Table 1, only results from the *Basic* or *Advanced* rewritten texts even lead to F1 scores lower than these majority class percentages⁴.

A strength of our method comes with the *advanced* setting, which shows in some cases *both* strong protection against adversaries and higher semantic similarity to the original texts than only once-rewritten texts. If one assumes cosine similarity (CS) to be an indicator of preserved utility, this becomes a quite interesting finding, in contrast to the belief that higher privacy always comes at the

⁴As majority class guessing will result in zero false negatives but many false positives.

Baseline F1		95.32											
Mechanism		DP-BART						DP-PROMPT					
ϵ		625			1875			137			206		
		F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow
Rewritten		25.72	50.12 _{0.8}	0.31	22.02	70.91 _{0.6}	0.57	17.92	18.84 _{0.7}	0.23	19.71	19.23 _{1.9}	0.44
Basic 2x		18.96	26.28_{0.1}	0.31	27.11	39.34_{1.0}	0.50	10.86	17.57_{0.0}	0.19	13.47	17.59 _{0.0}	0.41
Advanced 2x		25.20	33.62 _{0.4}	0.42	40.29	48.11 _{1.2}	0.53	9.42	17.57 _{0.0}	0.38	12.95	17.57 _{0.0}	0.48

(a) Yelp Empirical Privacy Results.

Baseline F1		73.23											
Mechanism		DP-BART						DP-PROMPT					
ϵ		625			1875			137			206		
		F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow	F1 (stat.) \downarrow	F1 (adapt.) \downarrow	CS \uparrow
Rewritten		60.39	60.03 _{2.2}	0.36	59.38	65.04 _{0.1}	0.62	58.53	58.08_{0.0}	0.22	61.99	60.37 _{1.6}	0.43
Basic 2x		55.99	58.51_{0.7}	0.28	59.48	61.47 _{0.7}	0.51	54.49	58.10 _{0.0}	0.16	56.63	59.91_{1.3}	0.33
Advanced 2x		55.17	59.38 _{0.5}	0.42	60.90	60.88_{2.0}	0.58	49.61	58.09 _{0.0}	0.36	54.73	60.30 _{0.3}	0.45

(b) Trustpilot Empirical Privacy Results.

Table 1: Empirical Privacy Results for Yelp and Trustpilot. *Rewritten* denotes the DP rewritten texts, while *Basic 2x* and *Advanced 2x* denote the result of our proposed post-processing methods (basic and advanced users). *Baseline F1* denotes the adversarial performance on the original, non-privatized texts. For each experiment setting, the best scoring result is bolded. For the adaptive (adapt.) attacker setting, the reported score is an average of three training runs, and the standard deviation is given as a subscript. *CS* denotes the average cosine similarity score between original and rewritten text.

trade-off of lower utility. Even where the once-rewritten text scores the highest in terms of *CS*, the empirical privacy gains shown by the advanced double-rewritten text are much more significant than the loss in *CS*, e.g., DP-BART ($\epsilon = 625$) on the Yelp dataset.

6.2 Basic vs. Advanced

An interesting point of analysis is the comparison between our two proposed methods, namely the *basic* and *advanced* users. In short, one clear winner between the two methods is not directly discernible from the results, as each showcases particular strengths.

While the advanced rewritten results achieve the best score most often in terms of empirical privacy scores, the basic rewritten method still outperforms the singly rewritten text more often (6 vs. 3 times), making a case for the “simpler” method that does not require extra fine-tuning. Interestingly, the basic rewritten texts achieve the lowest *CS* scores in all tested scenarios.

The promise of the *advanced* user is clear, as discussed above, particularly in its ability to improve privacy and semantic similarity simultaneously. One must keep in mind, however, that this advanced method necessitates the presence of *domain-specific* data to fine-tune the rewriting model. Therefore, the choice between basic and advanced usage of our post-processing method is ultimately contingent upon available resources as well as user preference.

6.3 A Case for Rewriting Again

In the above analysis, we pose that the observed benefits of post-processing DP rewritten texts make the case for adopting our proposed pipeline in empirical privacy evaluations. This method, while incurring the cost of extra training on the user side, presents a clear incentive for users in the DP rewriting scenario: the DP guarantee is upheld, while also producing output texts with higher privacy and semantic similarity to the original texts. We also present a method

that is *mechanism-agnostic*, meaning that this post-processing method can be run following any DP rewriting process. Moreover, the *basic* scenario, which does not require domain-specific private data, enables the open-sourcing of the proposed post-processing models to allow for private fine-tuning (advanced user).

Beyond this, our analysis of the output texts reveals that we also begin to tackle some overarching challenges of DP text rewriting, namely in the readability and coherence of the output texts. As noted by previous works [19, 28], DP text rewriting, particularly at higher privacy levels, runs the risk of producing outputs that are incoherent or repetitive. This comes as a side effect of the random noise addition to text representations, an inevitable result in satisfying DP. In performing a post-processing step on top of DP rewritten texts, we aim to alleviate these challenges by producing better semantically aligned texts. This is ensured by the inherent capability of (large) language models to generate such texts.

To solidify this point, we present selected examples of text outputs in Appendix A. One can argue that the texts produced by both the *basic* and *advanced* methods appear to be more fluid and coherent, as compared to their singly DP rewritten counterparts.

6.4 Limitations and Further Considerations

The discussion of the merits of our proposed method is not complete without a discussion of its limitations, including those of our evaluation, as well as the potential drawbacks of certain use cases.

Looking to the second rewriting process itself, namely with either model *T* or *T++*, one may observe that the process is dependent on the fine-tuning of a given language model. In this work, we evaluate one particular model, namely FLAN-T5-LARGE, and therefore it remains a point for future work to investigate the effect of model choice (architecture) and model size (i.e., number of parameters) on the method that we describe in this work.

Beyond the choice of rewriting model, the curation of data for the task of post-processing DP rewritten outputs is also seemingly quite important. As shown by our results, the advanced model $T++$, in general, achieves higher CS scores than the basic setting, which can be attributed to the fact that the model was further fine-tuned on domain-specific data aligned to the target data. Even before this, the choice of public corpus for the creation of the *aligned public corpus* is also important, as this serves as the basis for both the basic and advanced setups. In this work, we choose the C4 corpus as a reasonable public corpus, but further studies may be well-served to expand this to other text corpora. In addition, the question of *how much* data should be used to fine-tune the models is also not explored in this work, as we simply choose a random 100k sample.

We design our post-processing method to be *mechanism-agnostic*, meaning it can be run on top of any DP rewritten text outputs, as long as the mechanism is known and implementable. Despite this fact, we hypothesize that the nature of a given DP rewriting mechanism can also play a significant role in the effectiveness of the double-privatized outputs. Looking to the results of evaluation, where we study two distinct mechanisms, one can already see this effect in action. In particular, the empirical privacy results of DP-PROMPT tend to be stronger than those of DP-BART, regardless of whether our method is applied or not. This is most plausibly explained by the manner by which each mechanism rewrites, where DP-PROMPT models the task as *paraphrasing*, which often results in much shorter output texts that already “compress” much of the information of the originals. In contrast, DP-BART often much better mirrors the original text length, offering more space for semantic closeness to the original text – this is reflected by the generally higher CS scores, although this is difficult to equate across different mechanisms with differing effective ϵ scales.

A potential concern with arising from the second rewriting of texts comes with the possibility for *loss in factuality*, which comes as a result of both the double rewriting process, as well as from the known ability of language models to *hallucinate* information. The effect of this generation of seemingly plausible, yet potentially factually incorrect, information is outside the scope of our work, but presents an interesting starting point for future works.

A final consideration involves the inevitable fact that our proposed method adds extra overhead to the process of DP text rewriting, a task that can require significant resources even when not using LLMs in the underlying mechanism [28]. This is especially the case when considering the task placed upon to user to fine-tune a model locally. Although this can be alleviated by open-sourcing the base model T , the task of (re)rewriting considerably adds to the overall time requirement of DP text rewriting. Nevertheless, given the results of our empirical evaluations, we pose that the benefits of this extra task can be weighed individually by each user.

7 RELATED WORK

The study of Differential Privacy in Natural Language Processing can be traced back to the creation of synthetic Term-Frequency Inverse Document Frequency (TF-IDF) vectors [35], yet the first work on DP in the rewriting scenario was proposed using a generalized form of DP called *metric* DP [10]. Since then, several works have been proposed to improve upon the notion of metric DP for NLP,

mainly in the study of different metric spaces and distance metrics for word embeddings [2, 4, 5, 12, 27, 37–40].

Works surveying the field of DP in NLP have raised several challenges to the successful integration of DP in NLP tasks [20, 24]. A more recent survey categorizes the body of work into the characteristics of DP mechanisms, making the major distinction between *gradient perturbation* and *embedding perturbation* methods [17]. In these works, the primary challenges of DP in NLP are highlighted, most notably balancing the privacy-utility trade-off, exploring the meaning of the ϵ parameter, formalizing what exactly is being protected under a given DP guarantee, and the transparent and reproducible evaluation of rewriting mechanisms [18].

In response to the challenge of producing semantically coherent DP rewritten outputs, recent works have shifted from word-level perturbations to higher syntactic levels, namely the document level. While older works focus on text generation with auto-encoder-type models [1, 21], more recent works have leveraged the generative capabilities of transformer-based language models [19, 24, 34]. Still other works focus on the sentence-level [26] or specifically on DP language modeling [33, 36, 43].

Despite the recent wealth of works on DP text rewriting, little to no solutions have been proposed to improve the utility and/or privacy preservation of the privatized texts *post-generation*. In particular, the property of *post-processing* as a potential benefit remains under-researched in NLP applications that integrate DP. It is here where our work is centered, with the goal of providing an intuitive post-processing step for realigning DP rewritten texts while also improving its privacy protection.

8 CONCLUSION

We propose a post-processing method for differentially private rewritten text, which aims to enhance both the empirical privacy and semantic similarity of rewritten text. The evaluation of our methods, both in the basic and advanced user settings, reveals that our proposed pipeline not only offers significant improvements in reducing adversarial advantage, but it is also successful in “realining” the rewritten texts to mirror the original texts more closely. An analysis of the results shows that it is indeed possible to increase both the privacy and utility (CS) of the rewritten texts, thereby presenting a viable method for post-processing DP rewritten texts with the goal of enhancing their strength against capable adversaries.

The limitations of our work as discussed in Section 6.4 provide a clear path forward for future research. Concretely, we propose the following studies to build on our work: (1) an investigation of the effect of different (L)LMs in their usage in our post-processing scenario, most notably testing model size and architecture, (2) a study of the extent to which our method works across DP rewriting mechanisms, particularly focusing on potential drawbacks and negative side effects, and (3) continuing research on novel methods for post-processing DP rewritten text with the goal of making these privatized texts both private and useful.

We see our work as an important step in improving the usability of Differential Privacy in NLP, which still largely remains an academic pursuit. By leveraging post-processing and language models *for good*, we hope that future works will follow suit in advancing data privacy in NLP while also ensuring its practical relevance.

ACKNOWLEDGMENTS

The authors would like to thank Alexandra Klymenko for her valuable contributions to this work.

REFERENCES

- [1] Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2019. ER-AE: Differentially Private Text Generation for Authorship Anonymization. In *North American Chapter of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:198147337>
- [2] Danushka Bollegala, Shuichi Otake, Tomoya Machide, and Ken-ichi Kawarabayashi. 2023. A Neighbourhood-Aware Differential Privacy Mechanism for Static Word Embeddings. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 65–79. <https://doi.org/10.18653/v1/2023.findings-ijcnlp.7>
- [3] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea). (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2280–2292. <https://doi.org/10.1145/3531146.3534642>
- [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] Hui Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jianyun Nie, Chengyu Wang, and Jamie Cui. 2022. A Customized Text Sanitization Mechanism with Differential Privacy. In *Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:258841508>
- [6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shang Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhry, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53. <http://jmlr.org/papers/v25/23-0870.html>
- [7] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at Scale: Local Differential Privacy in Practice. In *Proceedings of the 2018 International Conference on Management of Data* (Houston, TX, USA) (SIGMOD '18). Association for Computing Machinery, New York, NY, USA, 1655–1658. <https://doi.org/10.1145/3183713.3197390>
- [8] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12. [https://doi.org/10.1007/978-3-030-17138-4_6](https://doi.org/10.1007/11787006)
- [9] Kennedy Edemacu and Xiantao Wu. 2024. Privacy Preserving Prompt Engineering: A Survey. *arXiv preprint arXiv:2404.06001* (2024). <https://doi.org/10.48550/arXiv.2404.06001>
- [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] Wikimedia Foundation. [n. d.]. *Wikimedia Downloads*. <https://dumps.wikimedia.org>
- [14] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. <https://doi.org/10.48550/arXiv.2111.09543> [cs.CL]
- [15] 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=XPZlaoutsD>
- [16] 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>
- [17] 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>
- [18] 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 Hahn, 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>
- [19] 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>
- [20] 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 Miresghallah (Eds.). Association for Computational Linguistics, Seattle, United States, 1–11. <https://doi.org/10.18653/v1/2022.privatenlp-1.1>
- [21] Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Autoencoder based Differentially Private Text Transformation. In *Conference of the European Chapter of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:231750016>
- [22] 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>
- [23] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained Language Model for Text Generation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4492–4499. <https://doi.org/10.24963/ijcai.2021/612> Survey Track.
- [24] 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想找naacl.65>
- [25] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103. <https://doi.org/10.1109/FOCS.2007.66>
- [26] Casey Meehan, Khalil Mirini, 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>
- [27] Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024. 1-Diffractor: Efficient and Utility-Preserving Text Obfuscation Leveraging Word-Level Metric Differential Privacy. *arXiv preprint arXiv:2405.01678* (2024). <https://doi.org/10.48550/arXiv.2405.01678>
- [28] 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>
- [29] Joseph P. Near and Chiké Abuah. 2021. *Programming Differential Privacy*. Vol. 1. <https://programming-dp.com/>
- [30] 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>

- [31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. <http://jmlr.org/papers/v21/20-074.html>
- [32] 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*. Association for Computational Linguistics. <https://arxiv.org/abs/1908.10084>
- [33] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. 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*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2848–2859. <https://doi.org/10.18653/v1/2022.naacl-main.205>
- [34] 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>
- [35] 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* (Ann Arbor, MI, USA) (SIGIR ’18). Association for Computing Machinery, New York, NY, USA, 305–314. <https://doi.org/10.1145/3209978.3210008>
- [36] Xinwei Wu, Li Gong, and Deyi Xiong. 2022. Adaptive Differential Privacy for Language Model Training. In *Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022)*, Bill Yuchen Lin, Chaoyang He, Chulin Xie, Fatemehsadat Miresghallah, Ninareh Mehrabi, Tian Li, Mahdi Soltanolkotabi, and Xiang Ren (Eds.). Association for Computational Linguistics, Dublin, Ireland, 21–26. <https://doi.org/10.18653/v1/2022.fl4nlp-1.3>
- [37] 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>
- [38] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A Differentially Private Text Perturbation Method Using Regularized Mahalanobis Metric. In *Proceedings of the Second Workshop on Privacy in NLP*, 7–17. <https://doi.org/10.18653/v1/2020.privatenlp-1.2>
- [39] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a Utilitarian Approach to Privacy Preserving Text Generation. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, 11–20. <https://doi.org/10.18653/v1/2021.privatenlp-1.2>
- [40] 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>
- [41] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf
- [42] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A Survey of large language models. *arXiv preprint arXiv:2303.18223* (2023). <https://doi.org/10.48550/arXiv.2303.18223>
- [43] Xin Zhou, Yu Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. 2023. TextObfuscator: Making Pre-trained Language Model a Privacy Protector via Obfuscating Word Representations. 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, 5459–5473. <https://doi.org/10.18653/v1/2023.findings-acl.337>

A REWRITING EXAMPLES

Table 2 showcases rewritten examples from the Yelp dataset, specifically the eleventh row.

		Review
DP-BART ($\epsilon = 625$)	Original	Saget's has been picking up and dropping off my dry-cleaning and this service is pricey - but so worth it as a timesaver for me. John (the Manager) is terrific and customer service has been great - they bill me monthly and charge it right to my card. Best of all the cleaning is expertly done and my clothes look great.
	Rewritten	SOMEONE has been doing a lot of work for me, and I have been a customer of his for over a year and a half and this has been a huge loss for me. I have to be honest, this is the first time I've been doing this, and this is what I'm going to be doing for the last time.S has been picking up and dropping off
	Basic 2x	S has been picking up and dropping off your used motorcycle parts in North Kingstown NJ. Some folks may have noticed us recommending them via their facebook page but have never seen one of the vehicles they work on and would love to purchase so we contacted him and spoke with the owner. There are some good points about these cars (or bikes in general), but they seemed really nice. When I spoke with S, he suggested that I could order an entire bunch of new parts for each bike I wanted and that my current set-up was perfect. All of that makes me want to buy something from this guy...
	Advanced 2x	S has been picking up and dropping off food, and the service and promptness to put things away - on time. I really wanted this place to do better and ... well, my dad's truck has a faulty gear box that will run on all 3 batteries and this was the perfect solution to take advantage and fix it, but I guess my car is simply not going to stand by for two weeks.
DP-BART ($\epsilon = 1875$)	Rewritten	I have been using this service for years and it has been great for me. The service has been amazing. I have been picking up and dropping off my clothes at this location for years. I love the service and the way it is done. My favorite part is that the service is free of charge.The service is great and I love it.My favorite part of this service is that I have
	Basic 2x	I have been picking up and dropping off my clothes at the LVAD store since 2008 and it has always been great! They go above and beyond to make sure every customer is taken care of. My favorite thing about their service is that they are always in control of your belongings and if we need additional items they can be reordered, delivered or picked up on time. My only complaint is that sometimes there are times they change the clothing that they will pick and deliver, but it never changes my items.
	Advanced 2x	Awesome service and quality clothes. My favorite part is that they never have a wait time. Love being pick up and dropped off when you need it and the clean of the facility is always appreciated. The service has been great and the staff is super nice.
	Rewritten	local pug-whip
DP-PROMPT ($\epsilon = 137$)	Basic 2x	he Pugwop Club is happy to present the Brutus Muppets from New York. These cartoon Pugs have been in production since 1985! Their live debut was the New Century Fox TV cartoon series "Bump It Up". They also enjoyed several notable roles in television shows, including The People's Love Band, and are now on NBC. In addition to being part of new shows, including "Pugwit," and "Geek Squad" they were cast in movies such as Zombies, Monster High, Spy Kids, Wet'n Wild.
	Advanced 2x	I have had so many good meals here that I've always been reluctant to give them a solid review. My boyfriend and I shared a lovely bowl of Thai-Kogi Curry noodle with some curry chicken for lunch. It was an excellent start to our day. We chose rice as the base, which came topped with veggies from their organic veggie patch. This was made into a huge bowl along with plenty of crispy coconut, sweet basil leaves (whole root vegetables) and ginger in curry sauce. Loved it!
	Rewritten	"The Professional Dry Cleaners can help save time!"
	Basic 2x	We are happy to help you. "The Professional Dry Cleaners can help save time! The team is wonderful! They came on the day they needed us and saved me from an unpleasant experience.
DP-PROMPT ($\epsilon = 206$)	Advanced 2x	Good times, good service, and this is NOT a dry cleaner's. Why wash dirty sheets for me? They come at 10 minutes so there is usually nothing else they need to do, like run errands or pick up the kids in a diaper. When you walk in and walk through the doors you have a feeling that there's really someone there (or the front desk). Once inside, the store has a few rooms with nice linens (pricey if you ask for them) and some old school ambiance and everything you can hope for.

Table 2: Rewritten Examples from the Yelp Dataset. **Rewritten** denotes the base DP rewritten texts, while **Basic 2x** and **Advanced 2x** denote our basic and advanced post-processing methods, respectively.