Thinking Outside of the Differential Privacy Box:Case Study in Text Privatization with Language Model Prompting

Stephen Meisenbacher
Florian MatthesUniversity of Munichof Computation, Information and Technologyof Computer Science,
texttttstephen.meisenbacher,matthes@tum.de

Abstract

Large language models (LLMs) are increasingly used to process sensitive textual data, raising concerns about privacy leakage. Differential privacy (DP) offers strong theoretical guarantees but often at the cost of reduced utility, particularly in natural language processing tasks. In this paper, we explore an alternative approach to text privatization that leverages language model prompting rather than relying solely on traditional DP mechanisms. We introduce DP-PROMPT, a prompting-based framework designed to reduce personally identifiable information (PII) and author-specific signals in text while preserving semantic content and readability. Through extensive experiments on a blog dataset, we evaluate the privacy-utility trade-offs of our approach compared to established baselines. Our results demonstrate that carefully designed prompting strategies can effectively mitigate author identification risks and topic leakage while maintaining high semantic similarity and readability, suggesting that LLM-based rewriting can serve as a practical complement to differential privacy methods in text privatization tasks.

1 Introduction

The topic of privacy in Natural Language Processing has recently gained traction, which has only been fueled by the prominent rise of Large Language Models. In an effort to address concerns revolving around the protection of user data, the study of privacy-preserving NLP has presented a plethora of innovative solutions, all investigating in some form the optimization of the privacy-utility trade-off for the safe processing of textual data.

A well-studied solution comes with the integration of Differential Privacy (DP) (Dwork, 2006) into NLP techniques. Essentially, the use of DP entails the addition of calibrated noise to some stage in a pipeline, e.g., directly to the data or to model weights. This is performed with the ultimate goal of protecting the individual whose data is being used, aligned with the objective of Differential Privacy set out in its inception nearly 20 years ago.

The incentive of proving Differential Privacy is the mathematical guarantee of privacy protection that it offers, so long as its basic principles are adhered to. Particularly, important DP notions must be strictly defined, such as who the individual is, how data points are adjacent, and how data can be bounded. As such, the fusion of Differential Privacy and NLP introduces several challenges (Feyisetan et al., 2021; Habernal, 2021; Klymenko et al., 2022; Mattern et al., 2022). When generalized forms of DP are used or well-defined notions of DP concepts are lacking, the promise of DP becomes more of a shallow guarantee.

In this work, we critically view the pursuit of DP in NLP, focusing on the particular method of DP-PROMPT (Utpala et al., 2023). This method leverages generative Language Models to rewrite (paraphrase) texts with the help of a DP token selection method based on the Exponential Mechanism (Mattern et al., 2022). We run experiments on three rewriting settings: (1) DP, (2) Quasi-DP, and (3) Non-DP; the purpose of this trichotomy is to explore the benefits and shortcomings of DP in text rewriting. We define our research question as:

What is the benefit of integrating Differential Privacy into private text rewriting methods leveraging LMs, and what effect can be observed by relaxing this guarantee?

Our empirical findings show the advantages that incorporating DP into text rewriting mechanisms brings, notably higher semantic similarity and resemblance to the original texts, along with strong empirical privacy results. This, however, comes with the downside of generally lower quality text in terms of readability, particularly at stricter privacy budgets. These findings open the door to discussions regarding the practical distinction between DP and non-DP text privatization, where we present open questions and paths for future work.

The contributions of our work are as follows:

1. We explore the merits of DP in LM text rewriting through comparative experiments.
2. We evaluate DP-PROMPT in a series of utility and privacy tests, and analyze the difference in DP vs. non-DP privatization.
3. We call into question the merits of DP in NLP, presenting the benefits and limitations of doing so as opposed to non-DP privatization.

2 Related Work

Natural language can leak personal information (Brown et al., 2022) and it is possible to extract training data from Machine Learning models (Pan et al., 2020; Carlini et al., 2021; Mattern et al., 2023). In the global DP setting, user texts are collected at a central location and a model is trained using privacy-preserving optimization techniques (Ponomareva et al., 2022; Kerrigan et al., 2020) such as DP-SGD (Abadi et al., 2016). The primary drawback of this model is that user data must be collected at a central location, giving a data curator access to the entire data (Klymenko et al., 2022). To mitigate this, text can be obfuscated or rewritten locally in a DP manner before collecting it at a central location (Feyisetan et al., 2020; Igamberdiev and Habernal, 2023; Hu et al., 2024).

The earliest set of approaches of DP in NLP began at the word level (Weggenmann and Kerschbaum, 2018; Fernandes et al., 2019; Yue et al., 2021; Chen et al., 2023; Carvalho et al., 2023; Meisenbacher et al., 2024a), yet these methods do not consider contextual and grammatical information during privatization (Mattern et al., 2022; Meisenbacher et al., 2024c). Other works operate directly at the sentence level by either applying DP to embeddings (Meehan et al., 2022) or latent representations (Bo et al., 2021; Weggenmann et al., 2022; Igamberdiev and Habernal, 2023). DP text rewriting methods using generative LMs (Mattern et al., 2022; Utpala et al., 2023; Flemings and Annavaram, 2024) or encoder-only models (Meisenbacher et al., 2024b) have also been proposed.

3 Method

Here, we describe the base text privatization method that we utilize, as well as the variations which form the basis of our experiments.

3.1 DP-PROMPT

DP-PROMPT (Utpala et al., 2023) is a differentially private text rewriting method in which users generate privatized documents at the local level by prompting Language Models to rewrite input texts. In particular, the LMs are prompted to paraphrase a given text. The immediate advantage of this method comes with the flexibility in model choice as well as the generalizability to all general-purpose pre-trained (instruction-finetuned) LMs.

The integration of DP into this rewriting process comes at the generation step, where for each output token, a DP token selection mechanism is implemented in the form of temperature sampling. In Mattern et al. (2022), it is shown that the use of temperature can be equated to the Exponential Mechanism (McSherry and Talwar, 2007). Relating this mechanism to the privacy budget ε of DP, the authors show that $\varepsilon = \frac{2\Delta}{T}$, where T is the temperature and Δ is the sensitivity, or range, of the token logits. A fixed sensitivity can be ensured by clipping the logits to certain bounds.

For the purposes of this work, we perform all experiments using DP-PROMPT with the FLAN-T5-BASE model from Google (Chung et al., 2022).

3.2 Rewriting Approaches

Motivated by the DP-PROMPT rewriting mechanism, we introduce three privatization strategies based on its DP token selection mechanism:

1. **DP:** we use DP-PROMPT as originally introduced, namely by clipping logit values and scaling logits by temperatures calculated based on ε values. We test on the values $\varepsilon \in 25, 50, 100, 150, 250$. Logits are clipped based on an empirical measurement of logits in the FLAN-T5-BASE model[1].

2. **Quasi-DP:** we replicate the DP strategy without clipping, i.e., only using temperature sampling based on the abovementioned ε values. We call this quasi-DP since the temperature values T are calculated as if clipping was performed (i.e., sensitivity is bounded), but the unbounded logit range is actually used.

3. **Non-DP:** here, we do not use any clipping or temperature, but rather only vary the top-k parameter, or the number k of candidate tokens considered when sampling the next token. We choose $k \in 50, 25, 10, 5, 3$.

With these three privatization strategies, we aim to measure empirically the effect on utility and privacy by strictly enforcing DP, relaxing DP, and by performing privatization devoid of DP. In this way, one may be able to analyze the merits of DP-based text privatization methods, and furthermore, observe the theoretical guarantees of DP in action.

4 Experimental Setup and Results

As stated by Mattern et al. (2022), a practical text privatization mechanism should: (1) protect against deanonymization attacks, (2) preserve utility, and (3) keep the original semantics intact. As such, we design our experiments by leveraging multiple dimensions of a single dataset. The results of all described experiments can be found in Table 1.

4.1 Dataset

For all of our experiments, we utilize the Blog Authorship Corpus (Schler et al., 2006). This corpus contains nearly 700k blog post texts from roughly 19k unique authors. The corpus also lists the

[1] Specifically, to the range $(logit_mean, logit_mean + 4 \cdot logit_std) = (-19.23, 7.48)$, thus $\Delta = 26.71$.

ID, gender, and age of author for each blog post. Full details on the preparation of the corpus are found in Appendix A; pertinent details are outlined below.

We prepare two subsets of the corpus. The first, which we call *author10*, only considers blog posts from the top-10 most frequently occurring blog authors in the corpus. This subset results in a dataset of 15,070 blog posts spanning five categories.

The second subset, called *topic10*, is necessary as the classification of the gender and age attributes for the *author10* dataset would be a less diverse and challenging task. We first take a random 10

4.2 Utility Experiments

We perform utility experiments for both the *author10* and *topic10* datasets. To measure utility across all privatization strategies, we first privatize each dataset on all selected privatization parameters. As we choose 5 parameters (ε/T or k) for each of our three strategies, this results in 15 dataset variants, i.e., 15 results per metric, each of which represents the average between the two datasets.

Semantic Similarity

To measure the ability of each privatization strategy to preserve the semantic meaning of the original sentence, we employ two similarity metrics: BLEU (Papineni et al., 2002) and cosine similarity. Both metrics strive to capture the similarity between output (in this case privatized) text and a reference (original) text; BLEU relies on token overlap while cosine similarity between embeddings is more contextual.

We use SBERT (Reimers and Gurevych, 2019) to calculate the average cosine similarity (CS) between the original blog posts and their privatized counterparts. For this, we utilize three embedding models to account for model-specific differences: ALL-MINILM-L6-V2, ALL-MPNET-BASE-V2, and GTE-SMALL (Li et al., 2023). For each dataset, we report the mean of the average cosine similarity calculated for each model.

We also report the BLEU score between privatized texts and their original counterparts. This is done using the BLEU implementation made available by Hugging Face. As before, reported BLEU scores are the average across an entire dataset.

Readability

In addition, we also measure the quality and readability of the privatized outputs by using perplexity (PPL) (Weggenmann et al., 2022), specifically with GPT-2 (Radford et al., 2019).

4.3 Privacy Experiments

Using *author10* and *topic10*, we design three empirical privacy experiments, in which an adversarial classification model is trained to predict a sensitive attribute (authorship, gender, or age) based on the blog post text. For this, we fine-tune a DEBERTA-V3-BASE model (He et al., 2021) for three epochs, reporting the macro F1 of the adversarial classifier.

We evaluate the privatized datasets in two settings (Mattern et al., 2022; Weggenmann et al., 2022). In the static setting, the adversarial model is trained on the original training split and evaluated on the privatized validation split. In the more challenging adaptive setting, the adversarial classifier is trained on the private train split. Lower performance implies that a method has better protected the privacy of the texts. Note that the adaptive score represents the mean of three runs. For all cases, a random 90/10 train/val split with seed 42 is taken.

In addition to F1, we also report the relative gain metric (γ), following previous works (Mattern et al., 2022; Utpala et al., 2023). γ aims to capture the trade-off between utility loss and privacy gain, as compared to the baseline scores. For the utility portion of γ, we use the CS results. Baseline scores are represented by adversarial performance after training and testing on the non-private datasets. We report the γ with respect to the adaptive setting.

5 Discussion

In analyzing the results, we first discuss the merits of DP text privatization. At stricter privacy budgets (lower ε), only the original DP-PROMPT is able to present significant gains, as showcased with $\varepsilon = 25$. At these lower values, one can also observe the benefits of enforcing DP via logit clipping, which results in higher CS and BLEU retention while outputting generally more readable text (much lower PPL). This trend with PPL holds for all scenarios of DP vs. Quasi-DP, making a clear case for proper bounding in DP applications.

In studying DP vs. Quasi-DP further, we notice that the distinction between the two, particularly at higher ε values, becomes somewhat opaque. In fact, Quasi-DP outperforms DP in terms of empirical privacy in many of the higher privacy budget scenarios. This would imply that a DP mechanism leveraging temperature sampling only becomes effective and sensible with stricter privacy budgets.

An important point of comparison also comes with the study results of our Non-DP method. A strength of this method is highlighted by its ability at lower k values (analogous to less strict privacy budgets) to maintain high levels of semantic similarity (CS), while still achieving competitive empirical privacy scores. For example, in the case of $k = 3$, this method is able to outperform all $\varepsilon \geq 100$ for both DP and Quasi-DP. The BLEU scores for Non-DP would also imply that this method is better able to rewrite texts in a semantically similar, yet lexically different manner, as opposed to DP methods at high ε values (see Appendix D). These results make a case for Non-DP privatization in certain cases, and in parallel, provide a critical view of using DP at high ε values which lead to ineffective empirical privacy.

A final point that is crucial to discuss is grounded in the observed relative gains. Looking to the cumulative scores ($P\gamma$) of Table 1, one can notice that the only positive gains are observed at relatively low ε values, implying that only at these levels do the empirical privacy protections begin to outweigh the losses in utility. The utility scores in these cases, however, are quite difficult to justify in real-world scenarios, where semantic similarity is quite low and readability suffers greatly. These results in general showcase the harsh nature of the privacy-utility trade-off, where mitigating adversarial advantage often comes with less usable data.

6 Conclusion

Central to this work is the debate on the merits of Differential Privacy in NLP. To lead this discussion, we conduct a case study with the DP-PROMPT mechanism, juxtaposed with two “relaxed” variants. Our results show that while the theoretical guarantee of individual privacy may be important in some application settings, in others, it may become too restrictive to apply effectively. Conversely, the merits of DP may be observed in stricter privacy scenarios, where the need for tight guarantees does bring favorable privacy-utility trade-offs.

We call for further research in two directions: (1) rigorous studies on the theoretical and practical implications of DP vs non-DP privatization, and relatedly, (2) the continued design of privatization mechanisms outside the realm of Differential Privacy that aim to balance strong privacy protections with practical utility preservation. We hope that researchers may be able to harmonize the "best of both worlds", keeping in sight the need for practically usable privacy protection of text data.

Acknowledgments

The authors thank Alexandra Klymenko and Maulik Chevli for their contributions to this work.

Limitations

The foremost limitation of our work comes with the selection of a single base model for use with FLAN-T5-BASE. While further testing should be conducted on other (larger) models, we hold that our results can be generalized, since model choice was not central to our findings. Another limitation is the choice of ε (i.e., temperature) and k values, which were not selected in any rigorous manner, but rather based on the relative range of values presented in Utpala et al. (2023). The effect of parameter values outside of our selected ranges thus is not explored in this work.

Ethics Statement

An ethical consideration of note concerns our empirical privacy experiments, which leverage an existing dataset (Blog Authorship) not originally intended for adversarial classification. In performing these empirical experiments, the actions of a potential adversary were simulated, i.e., to leverage publicly accessible information for the creation of an adversarial model. As this dataset is already public, no harm was inflicted in the privacy experiments as part of this work. Moreover, the dataset is made up of pseudonyms (Author IDs) rather than PII, thus further reducing the potential for harm.

References

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

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

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, 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*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

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)*, pages 883–890. SIAM.

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*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, 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. 2022. Scaling instruction-finetuned language models. Preprint, arXiv:2210.11416.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

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*, pages 123–148. Springer International Publishing.

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*, volume 34.

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*, WSDM '20, page 178–186, New York, NY, USA. Association for Computing Machinery.

James Flemings and Murali Annavaram. 2024. Differentially private knowledge distillation via synthetic text generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12957–12968, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ivan Habernal. 2021. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. Preprint, arXiv:2111.09543.

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*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

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*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

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*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. Preprint, arXiv:2308.03281.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

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*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103.

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)*, pages 3367–3380, Dublin, Ireland. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024a. 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*, IWSPA '24, page 23–33, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024b. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9314–9328, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024c. 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)*, pages 174–185, Torino, Italia. ELRA and ICCL.

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)*, pages 1314–1331.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. CoRR, abs/1908.10084.

Jonathan Schler, Moshe Koppel, and Shlomo Argamon. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

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*, pages 8442–8457, Singapore. Association for Computational Linguistics.

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*, pages 305–314.

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

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*, pages 3853–3866, Online. Association for Computational Linguistics.

A A Blog Dataset Preparation

We outline the process of dataset preparation for the data used in this work. All prepared datasets are made available in our code repository.

We begin with the corpus made available by Schler et al. (2006), which contains 681,284 blog posts from 19,320 authors and across 40 topics. In particular, we use the version made publicly available on Hugging Face[2]](https://huggingface.co/datasets/tasksource/blog$_{a}uthorship_{c}orpus).Inthisversion, eachblo$

Next, noticing that out of all the "topics", many contained very few blogs, we only considered blogs with topics in the top 15 most frequently occurring topics. We also only consider blog posts with a maximum of 256 tokens, both for performance reasons and also to remove outliers (very long blog posts). These two steps resulted in a further filtered set of 162,584 blogs.

To prepare the *author10* dataset, we considered the 10 most frequently blogging authors in the filtered corpus. This translates to authors writing between 1001 and 2174 distinct blog posts, for a total of 15,070 blogs in the *author10* dataset.

To prepare the *topic10* dataset, we only consider blog posts from the filtered corpus which count in the top 10 most frequently occurring topics. Concretely, this consists of the following topics (from most to least frequent): Technology, Arts, Education, Communications-Media, Internet, Non-Profit, Engineering, Law, Science, and Government. With these topics, we take a 10

While the gender attribute is not altered in the *topic10* dataset, we bin the age attribute for a more reasonable classification task. We choose to create five bins from the age column, which ranges from the age of 13 to 48. Creating an even split between all age bins, we achieve the following bin ranges:

$(13.0, 23.0] < (23.0, 24.0] < (24.0, 26.0] < (26.0, 33.0] < (33.0, 48.0]$

Thus, the resulting *topic10* dataset contains 10 topics, 2 genders, and 5 age ranges.

[2] [https://huggingface.co/datasets/tasksource/blog$_{a}uthorship_{c}orpus$

B B DP-PROMPT Implementation Details

We implement DP-PROMPT by following the described method in the original paper (Utpala et al., 2023). As noted, we leverage the FLAN-T5-BASE model as the underlying LM.

To set the clipping bounds for our method, we run 100 randomly sampled texts from our dataset through the model and record all logit values. Then, we set the clipping range to $(logit_mean, logit_mean + 4 \cdot logit_std) = (-19.23, 7.48)$, as noted in the paper.

For the prompt template, we use the same simple template as used by Utpala et al. (2023), namely:

Document: [ORIGINAL TEXT] Paraphrase of Document:

As discussed in the original paper, we do not change the top-k parameter for DP-PROMPT in its output generation, both for the DP and Quasi-DP settings. This is left to the default Hugging Face parameter of $k = 50$.

Finally, for comparability, we limit the maximum generated tokens for all methods to 64.

For all privatization scenarios, we run DP-PROMPT (and its variants) on a NVIDIA RTX A6000 GPU, with an inference batch size of 16.

The source code for replication can be found at the following repository, which also includes our two prepared datasets used in the experiments:

https://github.com/sjmeis/DPNONDP

C C Training Parameters

For all training performed as part of our empirical privacy experiments, we utilize the Hugging Face Trainer library for model training. All training procedures use default Trainer parameters, except for a training batch size of 64 and validation batch size of 128. Dataset splits are always shuffled with a random seed of 42 prior to training or validation. All training is performed on a single NVIDIA RTX A6000 GPU.

D D Examples

Tables 2 and 3 provide rewriting examples for all tested parameters for a selected text sample from each of our two datasets.