Baseline F1	**95.32**											
Mechanism	DP-BART						DP-Prompt					
ε	625			1875			137			206		
	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑
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**	$\mathbf{26.28_{0.1}}$	0.31	27.11	$\mathbf{39.34_{1.0}}$	0.50	10.86	$\mathbf{17.57_{0.0}}$	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**	$\mathbf{17.57_{0.0}}$	**0.38**	**12.95**	$\mathbf{17.57_{0.0}}$	**0.48**

(a) Yelp Empirical Privacy Results.

Baseline F1	**73.23**											
Mechanism	DP-BART						DP-Prompt					
ε	625			1875			137			206		
	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑	F1 (stat.) ↓	F1 (adapt.) ↓	CS ↑
Rewritten	60.39	$60.03_{2.2}$	0.36	**59.38**	$65.04_{0.1}$	**0.62**	58.53	$\mathbf{58.08_{0.0}}$	0.22	61.99	$60.37_{1.6}$	0.43
Basic 2x	55.99	$\mathbf{58.51_{0.7}}$	0.28	59.48	$61.47_{0.7}$	0.51	54.49	$58.10_{0.0}$	0.16	56.63	$\mathbf{59.91_{1.3}}$	0.33
Advanced 2x	**55.17**	$59.38_{0.5}$	**0.42**	60.90	$\mathbf{60.88_{2.0}}$	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 ($\varepsilon = 625$) 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 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 "realigning" 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 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* (, 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 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. 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/10.1007/11787006

[9] Kennedy Edemacu and Xintao 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 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=XPZIaotutsD

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

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

[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