AI for Data Science: A Benchmark for Differentially Private Text Dataset Generators

Viktor Schlegel [1,2] **Yuping Wu** [2] **Warren Del Pinto** [2] **Goran Nenadic** [2] **Anil A Bharath** [1,3]

1. Introduction

In high-stakes domains such as healthcare, finance, and legal services, experts face mounting challenges in leveraging data effectively to support decision-making. For example, a study of electronic health records (EHR) spanning approximately 100 million patient encounters found clinicians spending over 16 minutes per encounter reviewing charts and documentation [1]. While recent advances in generative AI show promise in making sense of complex data with minimal context, their application in specialized domains remains challenging due to data compartmentalization (and therefore exclusion from pre-training corpora of generative AI models) for regulatory, privacy, and institutional reasons.

As such, the performance of data-driven algorithms often deteriorates in these specialized settings, exemplifying the "jagged frontier" of AI [2], where unpredictable model behaviour leads domain experts to hesitate in adopting these technologies despite their potential benefits. To overcome these challenges and enable practitioners to benefit from AI breakthroughs, there is a critical need to include such domain-specific data into the open domain and pre-training and evaluation corpora of state-of-the-art AI methods.

With the proliferation of generative AI specifically, the development of realistic synthetic datasets that capture the complexity and nuance of domain-specific data while maintaining strict privacy guarantees offers an intriguing solution to the problem. In this context, Differential Privacy (DP) [3] offers a potential solution by providing formal guarantees about the maximum influence any individual record can have on resulting synthetic data. This approach enables organizations to generate high-quality representative synthetic data to share with external AI practitioners, while maintaining provable privacy guarantees. However, the effectiveness of these techniques in specialized contexts requires rigorous benchmarking and evaluation – an area that remains underdeveloped.

In this extended abstract, we present our ongoing work concerned with creating a unified benchmark for text dataset generation under formal DP guarantees. We motivate the benchmark design, discuss preliminary findings and conclude with an outlook for future work and possible research avenues.

[1] *Imperial College London, Imperial Global Singapore* [2] *University of Manchester, Department of Computer Science, United Kingdom* [3] *Imperial College London, Department of Bioengineering, United Kingdom*. Correspondence to: Viktor Schlegel v.schlegel@imperial.ac.uk.

2. Benchmark Design

2.1 Challenges in Domain-Specific Benchmarking

The evaluation of current synthetic data generation approaches faces several critical challenges that limit their applicability to specialized domains:

Realism & Representativeness: Existing evaluations typically use general domain datasets like sentiment analysis corpora rather than domain-specific data. This bias toward public domain data excludes specialized fields that could benefit most from privacy-preserving data sharing methods. For example, a realistic benchmark for synthetic financial transactions or medical records would ideally include actual sensitive data as a "private" dataset, but this is challenging due to privacy concerns.

Privacy Budget Assumptions: Many approaches assume label distributions are publicly known, which becomes problematic in specialized domains where rare combinations of attributes can uniquely identify individuals: Data points with rare combinations of characteristics are most susceptible to privacy leakage [4]. Additionally, hyper-parameters for generation models are often optimized on private data without accounting for the privacy budget this consumes.

Empirical Privacy Verification: Many works either omit rigorous empirical evaluation of privacy leakage or substitute it with simplified checks. This is concerning because implementation errors can invalidate formal privacy guarantees [5], and the relationship between the privacy parameter ϵ and actual privacy leakage is complex—two systems with the same ϵ value may have substantially different vulnerability profiles.

2.2 Addressing the Challenges

Our benchmark design addresses these challenges through several key innovations:

For Realism & Representativeness: We propose using datasets with gated access mechanisms (requiring Data Usage Agreements) to ensure realistic benchmarking while maintaining privacy. This approach enables evaluation on actual sensitive domain data while preventing its complete public release. Additionally, we incorporate fully open foundation models [6] with transparent training corpora [7], enabling verification that benchmark "private" data hasn't been exposed during pre-training.

For Privacy Verification: We develop diagnostic datasets specifically designed for membership inference attacks (MIA), allowing cost-effective validation of implementation correctness without requiring generation of many synthetic datasets. While de-

tailed MIA results are part of ongoing work not reported in this abstract, this approach will provide critical verification of privacy claims.

Our benchmark evaluates both utility and fidelity while ensuring robust privacy guarantees. Utility assessment usually quantifies how useful synthetic data is for real downstream application tasks. We achieve this by training models on synthetic data and evaluating their performance on real data, measuring how well synthetic data supports tasks like document classification. Fidelity measures how well synthetic data captures statistical properties and patterns of the original domain-specific data using metrics like MAUVE [8], text length distributions, entity mentions, and lexical diversity.

3. Preliminary Results

To validate our benchmark design, we present preliminary results from our ongoing work in the healthcare domain. We conducted experiments using state-of-the-art differentially private text generation methods: AUG-PE [9] and DP-Generator [10]. We used three semi-publicly available healthcare datasets: HoC [11] (cancer hallmark identification in scientific literature), N2C2 2008 [12] (obesity and co-morbidity recognition in clinical discharge summaries), and PsyTAR [13] (adverse drug effect detection in social media posts).

Table 1: F1 scores (utility) for different approaches, downstream models, and privacy budgets (ϵ).

Downsteam Model	Method	$\epsilon = \infty$	$\epsilon = 4$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$
Hallmarks Of Cancer						
BERT-large	Original	71.9				
	DP-Gen	54.8	19.2	14.7	13.6	17.1
	AUG-PE	15.0	8.2	6.6	7.6	5.0
BioClinicalBERT	Original	68.7				
	DP-Gen	48.9	15.8	11.7	9.2	4.5
	AUG-PE	12.9	6.8	7.4	8.3	6.5
DeBERTa-xlarge	Original	52.2				
	DP-Gen	39.0	0.0	0.0	0.0	20.7
	AUG-PE	2.8	0.4	11.9	11.3	3.8
N2C2 2008						
Longformer-large	Original	87.7				
	DP-Gen	61.1	55.8	59.3	56.9	55.6
	AUG-PE	58.1	58.9	53.2	59.1	55.9
Clinical-BigBird	Original	71.6				
	DP-Gen	55.7	53.2	53.2	53.2	53.4
	AUG-PE	53.2	53.2	53.2	53.2	53.2
Clinical-Longformer	Original	60.1				
	DP-Gen	59.0	55.4	53.6	53.2	53.2
	AUG-PE	53.2	53.2	53.2	53.2	56.9
PsyTAR						
BERT-base	Original	79.7				
	DP-Gen	69.5	33.5	32.2	33.3	31.1
	AUG-PE	61.0	62.1	60.9	54.5	49.5
BERT-large	Original	80.4				
	DP-Gen	70.3	39.1	36.0	36.1	31.7
	AUG-PE	63.9	64.7	63.5	58.1	50.9
DeBERTa-xlarge	Original	82.1				
	DP-Gen	75.3	23.9	15.8	27.6	4.9
	AUG-PE	44.0	65.3	65.7	60.2	54.4

Our results, reported in Tables 1 and 2 reveal significant performance degradation compared to results reported on general domain datasets. Methods achieve only 34-58% of the performance of models trained on real data at reasonable privacy levels ($\epsilon \leq 4$). Strikingly, performance deterioration is already apparent without privacy guarantees ($\epsilon = \infty$), suggesting that underlying language models struggle with domain-specific complexities and that the scores reported on open-domain datasets indeed might be inflated due to benchmark leakage into pre-training data.

Fidelity measures show similar trends. MAUVE scores between synthetic and real healthcare data are near zero (comparable to scores between completely unrelated datasets), indicating substantial distribution differences. Text length distributions also fail to match the original data, with synthetic texts often showing markedly different length patterns than real domain-specific texts.

Table 2: MAUVE, entity and n-gram overlap scores (fidelity) for different approaches, downstream models, and privacy budgets (ϵ) across datasets.

Method	$\epsilon = \infty$ MAUVE↑ NER↓ n-gram 1|2|3↓	$\epsilon = 4$ MAUVE↑ NER↓ n-gram 1|2|3↓	$\epsilon = 2$ MAUVE↑ NER↓ n-gram 1|2|3↓	$\epsilon = 1$ MAUVE↑ NER↓ n-gram 1|2|3↓	$\epsilon = 0.5$ MAUVE↑ NER↓ n-gram 1|2|3↓
Hallmarks Of Cancer					
Original	0.994 1.043 0.63|2.50|3.84				
DP-Gen	0.011 2.598 1.88|4.49|5.50	0.011 2.994 1.79|4.38|5.57	0.012 3.067 1.88|4.49|5.50	0.011 2.491 1.91|4.49|5.67	0.011 2.598 1.99|4.67|5.65
AUG-PE	0.012 4.831 3.52|5.60|6.69	0.012 3.279 3.52|5.60|6.69	0.012 3.880 3.90|6.35|7.63	0.010 4.272 4.22|6.68|7.80	0.012 4.831 4.58|7.32|8.35
N2C2 2008					
Original	0.996 0.739 0.59|1.46|1.60				
DP-Gen	0.135 8.185 1.40|1.82|2.01	0.032 8.180 7.23|9.70|9.02	0.018 8.185 7.22|9.50|8.90	0.023 7.701 7.02|8.89|7.79	0.019 6.605 6.89|8.52|7.68
AUG-PE	0.017 7.664 8.15|10.31|9.82	0.017 8.849 6.72|9.72|10.02	0.019 8.735 7.05|9.69|9.04	0.019 7.784 7.48|10.27|10.12	0.017 8.858 8.15|10.31|9.82
PsyTAR					
Original	0.988 0.857 0.61|3.10|6.32				
DP-Gen	0.019 2.161 1.56|5.40|8.31	0.020 2.025 1.56|5.40|8.31	0.023 2.161 1.65|5.41|8.27	0.021 2.229 1.53|4.61|6.25	0.019 2.401 1.88|5.75|8.55
AUG-PE	0.017 5.397 4.54|9.19|11.62	0.020 4.808 3.41|7.81|9.56	0.017 5.066 3.91|7.90|9.69	0.017 5.515 4.23|8.17|9.61	0.017 5.397 4.54|9.19|11.62

Interestingly, entity-level and n-gram frequency divergences are less pronounced, suggesting that vocabulary is better preserved than discourse structure, coherence, or syntax in synthetic domain-specific texts.

4. Conclusion

We present a benchmark design for synthetic text generators with formal privacy guarantees together with preliminary empirical results on healthcare text data. They reveal significant challenges in generating high-quality synthetic data: current state-of-the-art methods face substantial performance degradation when applied to domain-specific datasets, even under weak privacy constraints, highlighting the limitations of approaches relying on foundation models pre-trained on general domain data. These findings underscore the critical need for domain-specific benchmarks that realistically represent specialized data without compromising privacy. Our work takes an important step toward creating standardized benchmarks that can accelerate progress in privacy-preserving synthetic data generation for high-stakes applications.

In future work, we will extend our benchmark in several key directions: *(i)* multimodal data generation to address complex relationships between different data types such as clinical text reports and medical images [14]; *(ii)* fully private data generation mechanisms that address unaccounted privacy leakage, particularly for datasets with rare attribute combinations; *(iii)* more sophisticated evaluation metrics covering lexical diversity, coherence, discourse structure, and human judgments; and *(iv)* stronger membership inference attacks based on changes in these metrics to better quantify real-world privacy risks.

Acknowledgements

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We are grateful for the support provided by Research IT in form of access to the Computational Shared Facility at The University of Manchester and the computational facilities at the Imperial College Research Computing Service[1].

References

[1] J Marc Overhage and David McCallie Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Annals of internal medicine*, 172(3):169–174, 2020.

[2] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*, 2023.

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

[4] Matthieu Meeus, Florent Guepin, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Achilles' Heels: Vulnerable Record Identification in Synthetic Data Publishing. In *Proceedings of the 28th European Symposium on Research in Computer Security, Part II*, pages 380–399. Springer, 6 2024.

[5] Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, and Emiliano De Cristofaro. The Elusive Pursuit of Replicating PATE-GAN: Benchmarking, Auditing, Debugging. *Forthcoming in Transactions on Machine Learning Research*, 6 2025.

[6] Dirk Groeneveld et al. OLMo: Accelerating the Science of Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:15789–15809, 2024.

[7] Luca Soldani et al. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, 2024.

[8] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, Zaid Harchaoui, and Paul G Allen. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828, 12 2021.

[9] Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54531–54560. PMLR, 3 2024.

[10] Xiang Yue, Huseyin A. Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:1321–1342, 2023.

[11] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan H"ogberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinform.*, 32(3):432–440, 2016.

[12] Ozlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 07 2009.

[13] Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, Hamideh Moayyed, and Somaieh Goudarzvand. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90, 2019.

[14] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a deidentified publicly available database of chest radiographs with free-text reports. *Scientific Data 2019 6:1*, 6(1):1–8, 12 2019.

[1] DOI: https://doi.org/10.14469/hpc/2232

ϵ	**Text**
	Original
–	*Over a 10 yr period I have gained 70 lbs Other than that, it helped me with my anxiety attacks and depression.*
	AUG-PE
∞	*Tireless life support systems crumble at my disposal as a high stakes gamble with my vitality wears thin every time, as concerns multiply into a formidable puzzle that wears me down with each wearisome day, until that miraculous pill started to work wonders on an escalating tide of anguish, easing the turbulent depths of anxiety*
4	*The initial relief from the medication has subsided for the last six months, and I started experiencing unbearable discomfort and lack of rest just one month into treatment. Or alternatively: Treatment started showing its effect just a month after that, which left me feeling irritable and struggling to get enough rest after just six short*
2	*The escalating inflammation in my gastrointestinal area caused persistent and worsening symptoms, rendering my previous choice of Zetricity's treatment ineffective. or As time passed, severe gastrointestinal symptoms worsen despite my initial treatment choice Zetrics, ultimately making its impact less significant. or My previous medication Zetics, chosen to*
1	*Anxious individuals often require treatment with caution and careful consideration, as medications may have varying degrees of efficacy, especially for certain conditions. However, personal experiences can diverge significantly, demonstrating that not all medications can lead to the desired results, even those aimed to address related symptoms. Various encounters can evoke emotional responses that may*
0.5	*"Sorry, I'm not familiar with the issue you're referring to, so I'd be happy to help, but I need more context or specific details for me to provide a more informed solution."*
	DP-GENERATOR
∞	*Totacking with many other people comments. Burns bouts of suicide. omicidalnt put away. headaches changes in line with omicinations. rage didn't seem to lessen. migraines. extreme sleeplessness. racing thoughts. shaky. nervous. decreasedensitivity. stomachache. difficulty to concentrate. lasting ones: obsessive obsess over medication. profuse sweating and dry mouth. tantrums. decreased emotion range. lack of energy. etc Weight gain. bad*
4	*No desire to control myself (Brobrewation before this meds, started to get on crowd). Liver extended, initially seemed okay, but woke up day after attempt with whole weekend bender, and with a half dose of nice red alcohol (Weeks before Brokenthrothing, back to Happy Drobie). Straight away after the dose was a complete failure, the alcohol is messing up his thinkin and myin as he was subject to ever tiny chore at the dispesn). Wanting to continue*
2	*During the first 48 hours of my life on this drug, I had no energy, joint aches and body aches, sleepiness, and nothing on a good appetite. Not a worry of drug effects. Was not the sleep I hoped for. Burning twice the muscle weight night. Viglaintness flush. Walter, I'm sleeping well and whatnot. Folds cheap and budget/riskaty fever was unnoticeable and kept. Unused controlled. Minimum of emotion, deathizhed*
1	*I could not remember those days in an excited manner unlike ever. Weight gain is the only ill effect(I am not used to madness, so this weirdness around dietism). That aside, on July 9th treadmill session with workouts - 2 aborted days. Went getting and walking uphill. Diarrhoeia really. But that was ok because I am used to it. Went out and back, checked all calorie sites. Headed towards sugar free sweet depressant diet. Can't find anything*
0.5	*Starting with Crestor and taking 200mg a day of Sunday's were great. One morning mind is not with me, so I had to take plus another of my two with 300mg starting on Mondays and 500mg from thnot to May. Poster il mio muta i come verificando tutti il bilanciamento intramondanei, ebbientesse inendingur deces a vuoi decenisti. Aggiugno ho dato 186mgz*

Fig. A1: Example generations from the PsyTAR dataset for the label "Adverse Drug Reaction".

Appendix A. Experiment Setup

Benchmarked Approaches: We benchmark two state-of-the-art approaches using `Meta-LLama-3.2-1B` as the generative model (using OLMo is planned for future work):

AUG-PE [9]: Generates differentially private synthetic text without model training, using only LLM API access. It iteratively selects texts most similar to private data and creates variations through paraphrasing, with privacy guaranteed by adding Gaussian noise to the comparison process.

DP-Generator [10]: Fine-tunes pre-trained language models with DP-SGD on private data, then uses the resulting model to generate synthetic text.

Datasets: We use three healthcare text datasets: HoC [11]: A multi-label classification dataset for cancer hallmarks in scientific abstracts, publicly available but domain-specific. n2c2 2008 [12]: Obesity and co-morbidity recognition in MIMIC-III discharge summaries, accessible via gated mechanism. PsyTAR [13]: Adverse drug effect detection in social media posts, accessible via gated mechanism.

Evaluation Protocol: We generate synthetic data from the training portions of these datasets, assuming label distributions are public (privatizing these will be addressed in future work). For multi-label data, we treat each unique label combination as a separate class.

For fidelity, we evaluate MAUVE scores and KL divergence of named entity distributions, text length distributions, and collocation divergences. For utility, we assess downstream classifier performance when trained on synthetic data and evaluated on real data. All measures are evaluated at privacy budgets $\epsilon \in \{\infty, 0.5, 1, 2, 4\}$.

Appendix B. Qualitative Examples

Figure A1 shows example generations of evaluated approaches at different levels of privacy guarantees. The deterioration becomes apparent, as DP-GENERATOR makes up new words (such as "Brobrewation") or switches language mid-example. AUG-PE generates high-quality examples that are ultimately not very similar to the original example in style.