SynthTextEval: Synthetic Text Data Generation and Evaluation for High-Stakes Domains

Krithika Ramesh[♠] **Daniel Smolyak**[♢] **Zihao Zhao**[♠]
Nupoor Gandhi[♣] **Ritu Agarwal**[♠] **Margrét Bjarnadóttir**[♢] **Anjalie Field**[♠]
[♠]Johns Hopkins University
[♢]University of Maryland, College Park [♣]Carnegie Mellon University
{kramesh3, zzhao71, ritu.agarwal, anjalief}@jhu.edu
{dsmolyak, mbjarnad}@umd.edu
nmgandhi@cs.cmu.edu

Abstract

We present SYNTHTEXTEVAL, a toolkit for conducting comprehensive evaluations of synthetic text. The fluency of large language model (LLM) outputs has made synthetic text potentially viable for numerous applications, such as reducing the risks of privacy violations in the development and deployment of AI systems in high-stakes domains. Realizing this potential, however, requires principled consistent evaluations of synthetic data across multiple dimensions: its utility in downstream systems, the fairness of these systems, the risk of privacy leakage, general distributional differences from the source text, and qualitative feedback from domain experts. SYNTHTEXTEVAL allows users to conduct evaluations along all of these dimensions over synthetic data that they upload or generate using the toolkit's generation module. While our toolkit can be run over any data, we highlight its functionality and effectiveness over datasets from two high-stakes domains: healthcare and law. By consolidating and standardizing evaluation metrics, we aim to improve the viability of synthetic text, and in turn, privacy-preservation in AI development.

1 Introduction

Currently, in contexts involving sensitive data like healthcare and social services, accessing domain-specific text to develop AI systems and assistive tools necessitates strict data-sharing agreements and well-defined protocols to reduce privacy risks and maintain compliance with regulatory standards (de Kok et al., 2023; Alberto et al., 2023). Even with careful protocols, privacy-preservation is a key challenge, particularly if systems that have been developed using sensitive data are accessible to the public. While text anonymization can be used to redact or replace private identifiers from text, there is still significant risk of sensitive attributes in the data being de-identified or exposed (Staab et al., 2024; Xin et al., 2024; Pang et al., 2024).

Synthetic data offers a possible alternative means of privacy-preservation. If synthetic text is sufficiently reflective of real text, it could facilitate the development of tools or conducting analyses, while minimizing the risk of a privacy breach. Beyond privacy-preservation, there has been a substantial amount of interest in other synthetic text applications, including fine-tuning task-specific models, serving as a proxy for human evaluations and auditing models (Mitra et al., 2023; Xu et al., 2023; Tan et al., 2024; Huang et al., 2023; Min et al., 2023). Although approaches for generating synthetic text may differ depending on the intended use-case, they typically share the same underlying objective - to produce diverse, high-quality data with patterns and signals that mimic the characteristics of real-world data distributions (Wu et al., 2024; Liu et al., 2024).

Despite substantial recent interest in privacy-preserving text generation (Yue et al., 2023; Mattern et al., 2022; Wu et al., 2023; Tang et al., 2024; Nahid and Hasan, 2024), evaluation criteria vary across studies, leading to inconsistencies in conclusions and results that overlook key aspects of data, such as fluency and privacy (Ramesh et al., 2024). Existing frameworks for assessing synthetic text quality either lack a wide range of criteria and metrics or do not provide easy-to-use tools to evaluate one's own data (Belgodere et al., 2024; Chim et al., 2024). Given the importance of synthetic data generation and its potential use in high-stakes domains, there is a critical need for a thorough and consistent evaluation approach that supports comparability across studies. To this end, we present SYNTHTEXTEVAL: an open-source toolkit[1] for evaluating synthetic text and comparing generation methods. An overview of the architecture is provided in Figure 1. While generalizable to many syn-

[1]GitHub Repository: https://github.com/kr-ramesh/synthtexteval

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 487–499
November 4-9, 2025

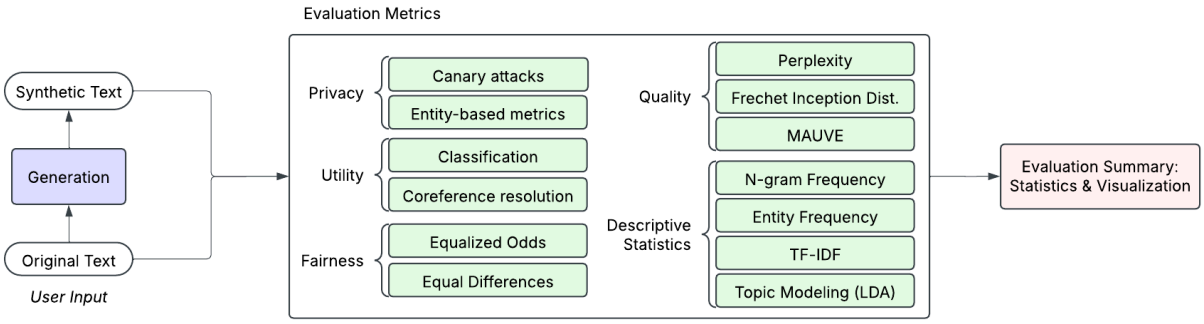

Figure 1: Architecture overview of SYNTHTEXTEVAL.

thetic text use cases, our toolkit particularly aims to enable the promising potential synthetic text has for facilitating privacy-preserving data sharing in high-stakes domains.

SYNTHTEXTEVAL has three core features that support this endeavor. First, it provides functionality to generate synthetic text with differentially private guarantees. Second, it implements a principled suite of generalizable metrics for evaluating the utility, privacy, fairness and quality of synthetic text, along with a descriptive outline of the data distribution. Third, a user-friendly GUI enables manually comparing synthetic text with real text, supporting the solicitation of qualitative feedback from non-technical domain experts. We further provide built-in support for generating and evaluating synthetic data in two domains with sensitive data: law (Pilán et al., 2022a) and healthcare (Johnson et al., 2016), but our toolkit is crucially data-agnostic, allowing users to run our generalizable metrics over their own datasets. Although our work is motivated by a clear need for standardized evaluation practices in high-stakes domains, it is broadly applicable to synthetic data for any task. By providing standardized data-agnostic metrics for evaluating synthetic data quality, we aim to assist researchers and practitioners in developing more robust and trustworthy AI systems and tools.

2 Background

Several factors influence the viability of synthetic data, including the faithfulness of the data to the source distribution, the diversity of the generated text, and its usability for developing downstream applications (Long et al., 2024). Prior work has often used evaluation criteria targeted to specific use cases, which make synthetic data generation methods difficult to compare. For instance, in generating synthetic text for privacy preservation, studies report metrics targeting training data leakage (Yue et al., 2023), but do not assess if the synthetic data introduces unfairness. In contrast, a system that aims to use synthetic data to address data imbalance may report results from groupwise performance metrics, but not privacy leakage (Pereira et al., 2024). Although not every metric is relevant for every use case, many of them do intersect: synthetic data is not a viable solution for privacy preservation if it introduces unfairness.

Even when the specified criterion is shared (e.g., privacy), studies often use different evaluation metrics , which can be contradictory and preclude comparability of methods (Friedler et al., 2021). Motivated by the need for consistent evaluation, our framework implements evaluation metrics that have been commonly used in prior work, but lack standardization (e.g., classification, canary attacks), as well as introduces metrics that have been shown to identify weakness in synthetic data (e.g., coreference resolution, fairness criteria) but are rarely reported (Ramesh et al., 2024).

2.1 Synthetic Data Evaluation Frameworks

Our toolkit differs from existing frameworks for evaluation of synthetic text in its wide range of criteria and metrics and provision of easy-to-use tools to evaluate one's own synthetic text data. Belgodere et al. (2024) provides a framework for auditing synthetic data across modalities, including tabular, time series, and text, and across fidelity, privacy, utility, and fairness metrics. However, due to their broad focus across modalities, their metrics are less tailored to text data, basing the evaluation on embeddings only. Chim et al. (2024) provides a text-specific auditing framework, focusing on user-generated text in messaging and social media contexts. They include metrics for downstream utility, privacy, and text fidelity, but do not include

	Evaluation Metrics					Ease-of-Use		Modality
	Utility	Privacy	Fairness	Quality	Descriptive	Code Package	GUI	Natural Language
SYNTHTEXTEVAL	✓	✓	✓	✓	✓	✓	✓	✓
Belgodere et al. (2024)	✓	✓	✓	✓	✓	✗	✗	✓
Chim et al. (2024)	✓	✓	✗	✓	✗	✗	✗	✓
Gehrmann et al. (2021)	✓	✗	✗	✓	✗	✓	✓	✓
Patki et al. (2016)	✓	✓	✓	✓	✓	✓	✓	✗
Qian et al. (2023)	✓	✓	✓	✓	✓	✓	✓	✗

Table 1: Comparison of evaluation frameworks for private synthetic text data generation.

descriptive statistics or fairness metrics. Neither framework provides a user-friendly code package nor a GUI to support qualitative evaluations.

In addition to synthetic text evaluations frameworks, other benchmarks, such as Gehrmann et al. (2021), focus on model capabilities (e.g., developed on synthetic data) in the context of natural language generation tasks, but are less focused on metrics specific to the generated text. Complementary to our work, Patel et al. (2024) developed the DataDreamer tool to standardize workflows for generating synthetic text data with LLMs, providing functionality for both finetuning generative models and training downstream models on generated synthetic data. However, they do not provide any functionality to assess the generated synthetic data. Patki et al. (2016) and Qian et al. (2023) both provide comprehensive evaluation frameworks and code packages for synthetic tabular data, but do not support evaluation of synthetic text.

3 Design and Implementation

SYNTHTEXTEVAL contains a text generation module and a set of evaluation modules, for each category of evaluation metrics, and a GUI as shown in Figure 1. Modules can be run individually to focus on specific metrics or together for a full audit.

The text generation model includes optional privacy-preserving guarantees for users who require them. It offers functionality for a descriptive analysis of any corpora, along with filtering functionality to remove low-quality synthetic text samples. In the evaluation module, the package includes tools to train and test downstream models on synthetic data, as well as automated evaluation of text quality, privacy and fairness metrics. The complete set of evaluation metrics are further explained in the following subsections.

3.1 GUI

Although we implement comprehensive and generalizable evaluation metrics, fully automated metrics can miss more nuanced aspects of synthetic text quality, especially ones that may vary across domains. For example, leakage of a name of a judge in a legal text may be acceptable, whereas leakage of a patient name in clinical text is not. To account for this, our framework contains a GUI that provides two key functionalities: i) for a given synthetic text, it displays the most real texts (determined through embedding similarity) ii) for a given synthetic text that contains a named entity, it displays real texts containing the entity. The GUI further supports writing free-form comments on notes, which can be saved and shared. The GUI is shown in Figure 2.

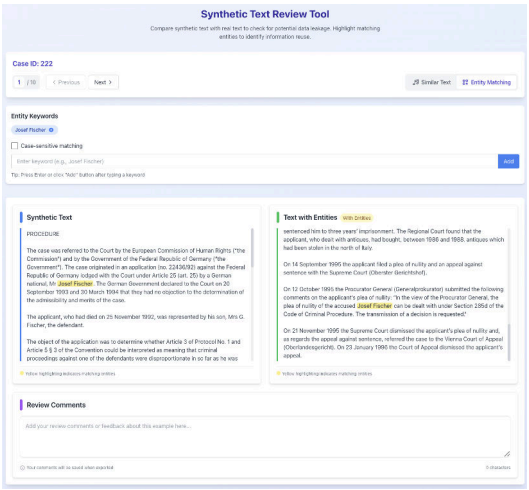

Figure 2: Our visual interface supporting exploration, comparison, and annotation of synthetic and real text.

This functionality in a user-friendly display allows non-technical domain experts to provide qualitative feedback on synthesized text, such as if synthesized text is factually accurate to real text and if it violates privacy by leaking information about entities.

3.2 Installation and Setup

SYNTHTEXTEVAL is an installable Python package[2], and the repository contains a number of user

[2] https://pypi.org/project/SynthTextEval/

resources including detailed setup instructions, information about various components, as well as test functions for each of the package's modules. In addition, we provide a demo Jupyter notebook and runnable experiment scripts for our case studies. A web-based version of our GUI is available for interactive exploration[3], and installation instructions for the desktop application are provided in the GitHub repository.

3.3 Downstream Utility

We assess the utility of synthetic data as a substitute for training models on real data with built-in support for two tasks: classification and coreference resolution. We select these two tasks both for their popularity in conducting evaluations (Table 1) driven by potential real-word use cases (Gandhi et al., 2023), and because the characteristics of the synthetic data that would increase its utility differs between the two tasks. For example, if the synthetic data successfully trains a high-performing classification model, this would suggest it contains a comprehensive set of features with similar frequency and contextual usage to the real data. However, lack of coherence or logical inconsistencies across synthesized documents may not be reflected in classification performance. In contrast, coreference resolution benefits from entity consistency, coherent structure, and the preservation of long-range dependencies.

In practice, a researcher or practitioner might use synthetic text as training data for these tasks by hiring annotators to label it, but collecting these "gold" annotations is too time-consuming and expensive to evaluate synthetic data quality across generation methods. Instead, for both tasks, we offer the functionality to generate "silver" annotations for the synthetic data by using an external pretrained model. New classification and coreference models can then be trained on this silver-annotated data, with the models' performance evaluated on a (real) gold standard test set. For classification, we focus on both F1 score and accuracy and for coreference resolution we focus on F1 score.

3.3.1 Fairness

SYNTHTEXTEVAL provides functionality to evaluate fairness in models trained on synthetic data. In domains such as healthcare and social services, the development of equitable systems is essential (Meng et al., 2022; Field et al., 2023). Prior research has identified differences in text data associated with difference groups, such as disparate rates of stigmatizing language in clinical notes (Harrigian et al., 2023), and examples where synthetic data increases downstream performance more for certain groups than others (Bhanot et al., 2021). Thus, for the tasks where we conduct utility evaluations, we additionally assess whether the distribution of utility is even across designated subgroups. We implement two commonly used fairness metrics: equalized odds (EO) and equality difference (ED). We provide full formulas in Appendix C.

3.4 Privacy

Our toolkit supports privacy evaluations for language models using both canary-based evaluations (Carlini et al., 2019) and entity-centric metrics (Ramesh et al., 2024). Out of proposed methods to quantify privacy leakage and memorization in LLMs (Carlini et al., 2022; Schwarzschild et al., 2024; Kim et al., 2023; Huang et al., 2024; Li et al., 2024), we focus specifically on measures that estimate the frequencies or likelihood of generation of spans of sensitive information and personal identifiable information (PII) in the synthetic text. These methods evaluate privacy risks in generated text, rather than vulnerability of models to attack (e.g., membership inference attacks), making them more suitable for evaluating synthetic text and also generalizable to diverse paradigms of privacy-preserving text generation, like sanitization approaches (Chen et al., 2023).

Canary-based evaluations measure memorization through the generation likelihood of injected canary phrases. While they offer a controlled way to assess data leakage (Yue et al., 2023), they are not always reflective of actual PII leakage (Ramesh et al., 2024). Thus, we additionally implement entity-centric evaluation criteria, which estimate the frequency of sensitive entities and their contexts in synthetic text. Appendix B provides additional details.

3.5 Text Quality

In order to evaluate the overall synthetic text quality, independently from specific downstream utility tasks, our toolkit implements a range of measures, including general perplexity-based evaluations as well as the reference-based metrics Fréchet Inception Distance (FID) (Heusel et al., 2017) and MAUVE (Pillutla et al., 2021). These metrics have

[3]Online interface: https://syntheticreview.cdhai.com/

been shown to often align with human judgment and can serve as valuable indicators of text quality, especially when human evaluation is impractical due to resource constraints. While these metrics have known 'blind spots' (He et al., 2023), using a combination of automated qualitative metrics can offer useful insight into the distributional similarity between synthetic and real data.

FID calculates the feature-wise mean and covariance matrices of embeddings for both synthetic and real text, and the Fréchet distance between these matrices serves as a measure of their similarity. MAUVE is a KL-divergence-based metric that compares the distributions of the synthetic and real text in the embedding space of a selected LLM. MAUVE's strength lies in its usefulness for relative comparisons regardless of embedding model choice, as it is robust to scaling, quantization, and embedding choices, while also correlating strongly with human judgment compared to other automated metrics. Perplexity assesses text quality by measuring the likelihood that a provided base LLM would generate a given text segment. A higher perplexity score indicates greater semantic coherence (Colla et al., 2022). Further details on how each metric is calculated are in Appendix D.

3.6 Descriptive Statistics

Users can conduct an exploratory text analysis of the synthetic data and make comparisons with real data by using the descriptive statistics module. Some of the functions included are an n-gram frequency analysis to detect common collocations and phrase patterns, TF-IDF computations to assess the relative importance of words across documents and an analysis of the most and least frequent entities in the text distributions. Additionally, to extract underlying thematic structures, the package employs Latent Dirichlet Allocation (LDA) for topic modeling, showing the user dominant topics and their associated keywords.

3.7 Generating synthetic text

Although our main focus is on facilitating evaluation, we additionally implement a popular synthetic data generation method, which involves fine-tuning an LLM with control codes, with or without differential privacy (Keskar et al., 2019; Yue et al., 2023; Ramesh et al., 2024). This functionality allows users to conduct an initial assessment of synthetic data quality in their own domains, as well as provides a baseline to compare against more customized generation methods.

Control codes are a special type of prefix in the input to the language model that contain labels corresponding to the features associated with the desired synthetic text (Keskar et al., 2019). These control codes are prepended to the prompt during training, allowing the model to learn the relationship between the labels and the corresponding output distributions of the synthesized text. We provide example control codes in the appendix E. During inference, users can supply their own control codes to the model, guiding it to generate text that matches the characteristics they desire for the intended use of the synthetic data.

We incorporate support for differentially private (DP) fine-tuning of models to generate synthetic text with privacy guarantees. DP techniques safeguard the participants in a dataset by making it difficult to infer whether any single individual's data was used. This is done by injecting carefully calibrated random noise to the results of any analysis, so that the outcome is nearly the same whether or not a given data point is included. For synthetic text generation, we achieve these privacy guarantees by fine-tuning our models with DP-SGD (Differentially Private Stochastic Gradient Descent) (Abadi et al., 2016). In DP-SGD, model gradients are clipped, and random noise is added during training, which limits the impact of any single data point and helps prevent the model from leaking sensitive information in its outputs. Further details about how DP works and its implementation in our study are provided in Appendix A.

While DP-based approaches offer well-defined upper bounds on the likelihood privacy leakage, this comes at the expense of model utility. Nevertheless, DP-generated synthetic text has been shown to be useful in reducing privacy risks while maintaining utility in certain downstream settings (Ramesh et al., 2024; Yue et al., 2023).

4 Validation and Example Usage

4.1 Experimental setup and objectives

Our first case study uses the Text Anonymization Benchmark (TAB), a dataset of 1,268 European court cases (Pilán et al., 2022b) extensively annotated to support evaluating text anonymization methods. We generate two synthetic datasets: one without DP ($\epsilon = \infty$) and one with DP ($\epsilon = 8$). We use the country and year of the court case as control codes for generation. Thus, our artificially created

classification task is to predict the country for each court case. A BERT$_{base}$ model is finetuned on each of the datasets. Our privacy evaluation focuses on the reoccurrence of personal information entities from the original dataset in the synthetic datasets.

Our second case study uses clinical notes from MIMIC-III (Johnson et al., 2016) and coreference-annotated clinical notes from the i2b2/VA Shared-Task (Uzuner et al., 2012). For the MIMIC-III data, we use the 10 most frequent International Classification of Diseases (ICD-9) codes as the control codes to generate synthetic data (similar to Al Aziz et al. (2021)), also with $\epsilon = \infty$ and $\epsilon = 8$.

To measure downstream utility, we evaluate performance on the multilabel and multiclass ICD-9 code prediction task for MIMIC-III, and the coreference resolution task for the i2b2/VA dataset. We compare performance for a BERT$_{base}$ model finetuned on only the original dataset to a model finetuned on the synthetic datasets. The fairness metrics for the ICD-9 code prediction task are calculated for both patient race and patient gender on the real clinical notes, while the entity-based privacy evaluations focus on the regurgitation of PII from the real data and the contexts in which they appear.

4.2 Results

We report classification utility for both datasets in Table 2 and remaining results in Appendix F. In general, results follow the trends we expect, where real text and non-DP synthetic text have the best utility and quality, but DP synthetic text has the best privacy. These results validate our toolkit's ability to reflect these properties.

4.2.1 TAB

Descriptive statistics show the synthetic data for TAB is on average about half as long and has half as many unique words (Table 3), although this is influenced by the user's choice of hyperparameters in generating the data. The synthetic data without DP ($D_{\epsilon=\infty}$) has higher Jaccard Similarity and Cosine Similarity to the real dataset compared to the synthetic data with DP ($D_{\epsilon=8}$). Regarding quality, both synthetic datasets are on par with the real test set for FID, but are lower quality according to the MAUVE and perplexity metrics (Table 4).

Utility results (Table 2) show the downstream model trained on $D_{\epsilon=\infty}$ has higher F1 and higher accuracy than the one trained on $D_{\epsilon=8}$. However, $D_{\epsilon=\infty}$ has a higher percent of PII leaked from the real data, indicating a higher privacy risk (Table 5). There is relatively high privacy risks for both synthetic datasets, likely due to the small size of the original TAB dataset. The $D_{\epsilon=8}$ dataset trades off slightly better privacy metrics for lower quality and utility compared to $D_{\epsilon=\infty}$.

		F1 Score	Accuracy
TAB	$D_{\epsilon=\infty}$	0.96 ± 0.035	0.99 ± 0.012
	$D_{\epsilon=8}$	0.54 ± 0.0 ↓-0.42	0.95 ± 0.0 ↓-0.04
Healthcare	D_{real}	0.67 ± 0.012	0.32 ± 0.016
	$D_{\epsilon=\infty}$	0.61 ± 0.003 ↓-0.06	0.26 ± 0.004 ↓-0.06
	$D_{\epsilon=8}$	0.40 ± 0.004 ↓-0.27	0.18 ± 0.005 ↓-0.14

Table 2: **TAB**: Difference in performance between models trained on data generated with DP and models trained on data generated without DP over TAB classification. **Healthcare**: difference in performance between real and synthetic data as training data for the ICD-9 classification task.

4.2.2 Healthcare

For both coreference resolution and mention detection, training on either synthetic dataset results in significantly lower utility compared to the real data, with $D_{\epsilon=\infty}$ and $D_{\epsilon=8}$ providing nearly identical performance (Table 6). In contrast, for the ICD-9 classification task (Table 2), there is a more pronounced drop in utility when using the differentially private synthetic data ($D_{\epsilon=8}$), while $D_{\epsilon=\infty}$ achieves higher performance than $D_{\epsilon=8}$ but still does not reach the performance of the model trained on the original dataset.

Fairness metrics (Table 7) suggest the model trained on the real data is most fair for race as the sensitive attribute, with the lowest FNED and EO metrics. $D_{\epsilon=\infty}$ is second highest, and $D_{\epsilon=8}$ has the worst fairness. For gender as the sensitive attribute, the models are all approximately equally fair across training datasets. On the privacy criteria, $D_{\epsilon=\infty}$ includes a higher percent of reproduced entities compared to $D_{\epsilon=8}$. Overall, our clinical notes evaluation shows that generating data without DP provides higher downstream utility and fairer downstream models at the cost of slightly higher privacy risks, compared to synthetic clinical notes generated with DP.

Conclusion SYNTHTEXTEVAL provides a toolkit to comprehensively evaluate synthetic text data across a wide range of metrics. This toolkit allows users the opportunity to assess the relative merits of synthetic data generation approaches, through the standardization of evaluation metrics, allowing for greater confidence in synthetic data

quality, especially in high-stakes domains where quality synthetic data is in high demand.

5 Ethics Statement

SYNTHTEXTEVAL is designed to lower barriers to development of AI systems in sensitive domains through comprehensive synthetic text data evaluation. Rather than rely on ad-hoc evaluation and comparison of synthetic text data, we hope our toolkit will help standardize comparisons and encourage responsible synthetic text data generation. In particular, our fairness and privacy metrics emphasize that high performance on utility and quality metrics are not on their own sufficient criteria for determination of worthiness for synthetic data publication and distribution.

However, while we have aimed to include wide range of metrics across many criteria, we caution that other context-dependent assessment may be necessary.An unintended consequence of this package could be a false sense of security from satisfactory performance on the included metrics, leading to less consideration of other context-specific evaluations. Additionally, there may be contexts where it is unethical or harmful to generate synthetic text data, regardless of performance on evaluations. We are committed to continuing to improve this package. Future efforts will focus on widening the evaluation criteria and metrics included, providing further visual or interactive tools for comparison, and enhancing reliability of the package.

Acknowledgments

We thank reviewers for their helpful feedback. This work was supported in part by the AI2050 Fellowship program by Schmidt Sciences. We also thank Tianli Xu for his valuable contribution in developing the annotation interface for this paper.

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. ACM.

Md Momin Al Aziz, Tanbir Ahmed, Tasnia Faequa, Xiaoqian Jiang, Yiyu Yao, and Noman Mohammed. 2021. Differentially private medical texts generation using generative neural networks. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–27.

Isabelle Rose I Alberto, Nicole Rose I Alberto, Arnab K Ghosh, Bhav Jain, Shruti Jayakumar, Nicole Martinez-Martin, Ned McCague, Dana Moukheiber, Lama Moukheiber, Mira Moukheiber, et al. 2023. The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet Digital Health*, 5(5):e288–e294.

Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. 2024. Auditing and generating synthetic data with controllable trust trade-offs. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*.

Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. 2021. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv:2202.07646*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

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.

Jenny Chim, Julia Ive, and Maria Liakata. 2024. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, pages 1–44.

Davide Colla, Matteo Delsanto, Marco Agosto, Benedetto Vitiello, and Daniele P Radicioni. 2022. Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134:102393.

Jip WTM de Kok, Miguel Á Armengol de la Hoz, Ymke de Jong, Véronique Brokke, Paul WG Elbers, Patrick Thoral, Alejandro Castillejo, Tomás Trenor, Jose M Castellano, Alberto E Bronchalo, et al. 2023. A guide to sharing open healthcare data under the general data protection regulation. *Scientific data*, 10(1):404.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in nlp tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1479–1492, New York, NY, USA. Association for Computing Machinery.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Nupoor Gandhi, Anjalie Field, and Emma Strubell. 2023. Annotating mentions alone enables efficient domain adaptation for coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10543–10558, Toronto, Canada. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. 2021. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, pages 11631–11642. Curran Associates, Inc.

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. 2023. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA. Curran Associates Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Wei Huang, Yinggui Wang, and Cen Chen. 2024. Privacy evaluation benchmarks for NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2615–2636, Miami, Florida, USA. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024. PrivLM-bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 54–73, Bangkok, Thailand. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *Preprint*, arXiv:2311.11045.

Md Mahadi Hasan Nahid and Sadid Bin Hasan. 2024. Safesynthdp: Leveraging large language models for privacy-preserving synthetic data generation using differential privacy. *Preprint*, arXiv:2412.20641.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, Minhui Xue, and Bo Li. 2024. Reconstruction of differentially private text sanitization via large language models. *Preprint*, arXiv:2410.12443.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024. Datadreamer: A tool for synthetic data generation and reproducible llm workflows. *arXiv preprint arXiv:2402.10379*.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

Mayana Pereira, Meghana Kshirsagar, Sumit Mukherjee, Rahul Dodhia, Juan Lavista Ferres, and Rafael de Sousa. 2024. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *Plos one*, 19(2):e0297271.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022a. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022b. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. 2023. Synthcity: facilitating innovative use cases of synthetic data in different data modalities.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. 2024. Rethinking LLM memorization through the lens of adversarial compression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and Lichao Sun. 2024. Unigen: A unified framework for textual dataset generation using large language models. *Preprint*, arXiv:2406.18966.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. 2023. Privacy-preserving in-context learning for large language models. *Preprint*, arXiv:2305.01639.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

A Background: Differential Privacy

Differential privacy (DP) offers a formal privacy guarantee that ensures that any individual's data cannot be inferred from a query applied to a dataset (Dwork et al., 2006, 2014). In other words, the result of such a query is nearly indistinguishable from the result of the same query applied to a dataset that either includes a modified version of the individual's data or excludes the record entirely, thereby preserving the individual's privacy. In this case, the notion of adjacency is defined as a difference of a single record in the original dataset D and the modified dataset D'. Formally, DP is defined as follows:

Definition: Given a dataset D and an adjacent dataset D', which is produced by removing or modifying a single record from D, a randomized algorithm $F : D \to Y$ is (ϵ, δ)-private if for any two neighboring datasets D, D', with the constraints $\epsilon > 0$ and $\delta \in [0, 1]$, the following holds true for all sets $y \subseteq Y$:

$$\Pr[F(D) \in y] \leq e^{\epsilon} \Pr[F(D') \in y] + \delta$$

The value of ϵ denotes the privacy budget, while δ specifies the likelihood that the privacy guarantee may fail. When δ is set to 0, this implies a purely differentially private setting with no probability of the guarantee being broken. The value of ϵ constrains how similar the outputs of both distributions are; a higher ϵ value indicates a greater privacy budget, meaning the algorithm is less private. DP guarantees that even if an adversary has access to any side-knowledge, the privacy leakage of (ϵ, δ)-DP algorithms will not increase. Additionally, another property of DP is that it ensures that any post-processing on the outputs of (ϵ, δ)-differentially private algorithms will remain (ϵ, δ)-differentially private.

We use DP-SGD (Abadi et al., 2016), a modification to the stochastic gradient descent (SGD) algorithm, which is typically used to train neural networks. DP-SGD clips the gradients to limit the contribution of individual samples from the training data and subsequently adds noise from a predefined type of distribution (such as a Gaussian distribution) to the sum of the clipped gradients across all samples. DP-SGD thus provides a differentially private guarantee to obfuscate the gradient update, thereby ensuring that the contribution of any given sample in the training data is indistinguishable due to the aforementioned post-processing property. This process ensures (ϵ, δ)-differential privacy for each model update. Given a privacy budget, number of epochs, and other training parameters, we can estimate the privacy parameters using estimation algorithms (Gopi et al., 2021).

B Privacy metrics

Canary-based evaluations (Carlini et al., 2019) involving crafting canaries that contain sensitive information and injecting them into the training data. The probability of generating the canary is estimated in contrast to the model trained on the corpus without the canary (a model that has "memorized" the canary would have a higher likelihood of generating it). In addition to this, we implement entity-centric evaluation criteria to quantify the Entity Leakage Percentage (ELP) and the proportion of an entity's context leaked, where the user can define the context window length (Ramesh et al., 2024).

The ELP is calculated as follows, where E_{leaked} is the number of entities leaked in the synthetic generations, and E_{total} represents the total number of entities in the training data.

$$P_{\text{entities}} = \frac{E_{\text{leaked}}}{E_{\text{total}}} \times 100$$

Similarly, we quantify the memorized spans of context where the entities appear as follows:

$$P_{\text{context}}(k) = \frac{C_{\text{leaked}}(k)}{C_{\text{total}}(k)} \times 100$$

where $C_{\text{leaked}}(k)$ is the number of occurrences where the entity and its surrounding context (of length k) are leaked in the synthetic generation and $C_{\text{total}}(k)$ is the total number of occurrences where the entity and its context (of length k) appear in the original training data.

C Fairness metrics

Where D is a set consisting of all subgroups corresponding to a demographic attribute within the dataset, Equalized Odds is defined as:

$$\text{EO}_D = \max\Big(\max_{d \in D}(\text{TPR}_d) - \min_{d \in D}(\text{TPR}_d), \max_{d \in D}(\text{FPR}_d) - \min_{d \in D}(\text{FPR}_d)\Big)$$

Equalized difference sums the differences between subgroup performance and overall performance for a specific metric. For example, FPR is

an important metric in domains such as healthcare, where false positive diagnoses can place increased burdens on patients in the form of unnecessary treatment or exams (Rajkomar et al., 2018). Thus, the False Positive Equality Difference (FPED) metric is the sum of the differences between the overall FPR for the entire dataset and the FPR for each subgroup:

$$\text{FPED} = \sum_{d=1}^{D} |\text{FPR}_{\text{overall}} - \text{FPR}_d| \quad (1)$$

ED metrics for true positive (TPED), true negative (TNED) and false negative (FNED) are computed similarly. Lower values of EO and ED scores indicate that the model's performance is more consistent across different subgroups.

D Automated Text Quality Evaluation Metrics

We provide the full equations to calculate our two automated qualitative evaluation metrics, MAUVE and Fréchet Inception Distance.

MAUVE is computed by first sampling human text $x_i \sim P$ (reference) and generated text $x'_i \sim Q$ (synthetic). Embeddings for this text are then obtained from an external model M (we use GPT-2 ()). These embeddings are quantized through clustering, yielding "low-dimensional discrete distributions that approximate each high-dimensional text distribution" (Pillutla et al., 2021). Finally, a divergence curve is calculated by taking the KL-divergence between the reference distribution and synthetic distribution as a mixture weight λ is varied, as shown in Equation 2, and MAUVE is calculated as the area under this curve.

$$C(P,Q) = \Big\{ \big(e^{-c\text{KL}(Q\|R_\lambda)}, e^{-c\text{KL}(P\|R_\lambda)}\big) : R_\lambda = \lambda P + (1-\lambda)Q,\ \lambda \in (0,1) \Big\}. \quad (2)$$

The Fréchet Inception Distance is obtained by measuring the Wasserstein-2 distance between the Gaussian obtained from the reference data (m, C) and the Gaussian of the synthetic data (m_w, C_w):

$$d^2((m,C),(m_w,C_w)) = \|m - m_w\|_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}$$

E Control Codes

We adapt the dominant approach from prior work (Yue et al., 2023): which involves first fine-tuning a pre-trained autoregressive language model on real, in-domain data and then generating synthetic data from the fine-tuned model. We compare fine-tuning the model with and without DP, where we use DP-SGD for differentially private fine-tuning. After fine-tuning, we utilize top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020) to generate diverse synthetic notes.

We condition the text generation on *control codes* (Keskar et al., 2019). During training, we prepend one or more labels associated with the text to the model input. We similarly prepend control codes during inference, where we sample the provided codes from their distribution in the training data. Thus, during training and inference, the probability distribution of the subsequent text $x = \{x_1, x_2 ... x_n\}$ is conditioned on the control code information c, which is described by the following equation:

$$P(x|c) = \prod_{i=1}^{n} P(x_i|x_1 ... x_{i-1}, c) \quad (3)$$

The format of the control code for the MIMIC-III data is as follows: *Long_Title: <diagnoses>, ICD9_CODE: <codes>, Gender: <gender>, Ethnicity: <ethnicity>*, where the <diagnoses> variable represents the long title form of the ICD-9 codes, information that is already provided with the MIMIC-III dataset. An example control code is as follows: *Diagnosis: Long_Title: Unspecified essential hypertension, Atrial fibrillation ICD9_CODE: 4019, 42731 Gender: Female Ethnicity: WHITE*

F Sample Usage Results

Domain	Avg Length Of Dsynth	Avg. Num Unique Words	Min Text Length	Max Text Length	Jaccard Similarity	Cosine Similarity
TAB (DP - inf)	617.725	265.882	452	745	0.185	0.522
TAB (DP - 8)	613.75	245.95	8	787	0.147	0.474
TAB (Real)	1342.42	487.966	185	5144	-	-

Table 3: Descriptive statistics for TAB across real and synthetic datasets.

	FID	MAUVE	Avg. Perplexity
$D_{real(test)}$	0.788	0.941	15.38
$D_{\epsilon=\infty}$	0.797	0.485	11.474
$D_{\epsilon=8}$	0.790	0.831	10.209

Table 4: Text quality metrics for TAB. The first column indicates the comparison dataset, with $D_{real(train)}$ as the reference dataset for each.

	Healthcare	TAB
	% Entities	**% Entities**
$D_{\epsilon=\infty}$	4.8	50.3
$D_{\epsilon=8}$	3.8	48.5

Table 5: Percent of unique PII from the training data that appear in the synthetic generations.

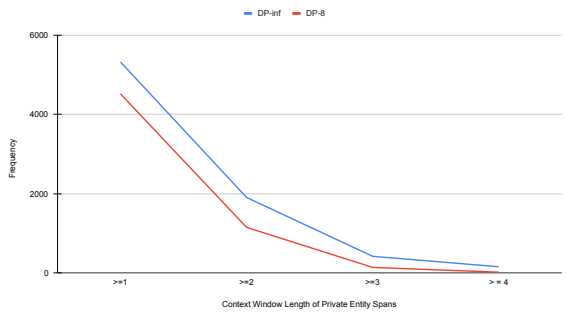

Figure 3: Memorization of private entities in the TAB dataset as context window length increases.

	Mention Detection	Coreference
$D_{real(gold)}$	0.800 ± 0.005	0.706 ± 0.006
$D_{real(silver)}$	0.659 ± 0.121	0.552 ± 0.126
$D_{\epsilon=\infty}$	0.650 ± 0.008	0.483 ± 0.002
$D_{\epsilon=8}$	0.656 ± 0.006	0.491 ± 0.003

Table 6: F1 scores for coreference and mention detection for MIMIC-III over entities from human-annotated test splits of the i2b2/VA datasets. All synthetic datasets are annotated with silver labels.

		FNED	Equalized Odds
Race	D_{real}	0.35 ± 0.0	0.20 ± 0.0
	$D_{\epsilon=\infty}$	0.39 ± 0.005	0.23 ± 0.001
	$D_{\epsilon=8}$	0.53 ± 0.014	0.30 ± 0.003
Gender	D_{real}	0.04 ± 0.0	0.04 ± 0.0
	$D_{\epsilon=\infty}$	0.02 ± 0.007	0.02 ± 0.007
	$D_{\epsilon=8}$	0.04 ± 0.005	0.04 ± 0.005

Table 7: Fairness evaluation for the MIMIC-III ICD-9$_{n=10}$ task, for the gender and race categories. Higher values indicate poorer group fairness performance.

	Rank ($\epsilon = \infty$ / 8)	Perplexity ($\epsilon = \infty$ / 8)
100 Insertions		
Name	4628 / 3356	35.80 / 53.19
Address	5 / 3967	16.52 / 61.37
Number	1 / 818	5.77 / 14.46
Email	1 / 1410	10.50 / 70.26
10 Insertions		
Name	5986 / 3378	49.72 / 54.10
Address	2276 / 4075	43.59 / 62.66
Number	902 / 841	9.43 / 14.61
Email	711 / 1452	37.81 / 72.08
1 Insertion		
Name	6037 / 3383	52.06 / 54.20
Address	3348 / 4081	54.50 / 62.79
Number	1084 / 838	9.82 / 14.63
Email	1941 / 1457	43.90 / 72.28
0 Insertions		
Name	6086 / 5265	44.81 / 57.54
Address	4565 / 3869	75.79 / 61.68
Number	1217 / 1522	11.80 / 13.33
Email	1003 / 3174	43.80 / 55.19

Table 8: Rank and perplexity metrics for canary attacks over Healthcare (MIMIC) data. Each entry is shown as $\epsilon = \infty$ / 8. Differential privacy ($\epsilon = 8$) increases both metrics, improving privacy for all canaries.