Generating and Evaluating Synthetic Data for Privacy Preservation in High-Stakes Domains

Anonymous ACL submission

Abstract

The difficulty of anonymizing text data hinders the development and deployment of NLP in high-stakes domains that involve private data, such as healthcare and social services. Poorly anonymized sensitive data cannot be easily shared with annotators or external researchers, nor can it be used to train public models. In this work, we develop methods to generate and evaluate synthetic data to facilitate the development of NLP in these domains without compromising privacy. We use language models fine-tuned with differential privacy to generate data and incorporate NLI-based filtering to improve text coherence. In contrast to prior work, we generate and evaluate data for fine-grained applications in real high-stakes domains. Our results show that prior simplistic evaluations have failed to highlight utility, privacy, and fairness issues in the synthetic data generated, and while NLI-based filtering can help alleviate some of these weaknesses, the quality of the synthetic data generated still necessitates further improvements.

1 Introduction

Widespread availability of public digitized text has greatly facilitated the advancement of natural language processing (NLP). Text processing could also be extremely valuable for processing high-stakes private data, like healthcare records (Panchbhai and Pathak, 2022), social workers' notes (Gandhi et al., 2023), or legal documents (Zhong et al., 2020). However, the need to maintain data privacy hinders the responsible development and deployment of models in these domains.

Building NLP often requires sharing data externally with contractors or researchers, as agencies like child protective services typically do not have in-house AI expertise. While this has been accomplished through data use agreements with individual teams, it still requires increasing the number of people who have access to sensitive data. Furthermore, limited sharing precludes the development of public benchmarks, which have proved crucial for standardizing model development, and any models trained on private data must themselves be treated as private, as NLP models are prone to outputting sensitive information from training data (Carlini et al., 2023). The risks of sharing data further lead to trade-offs between privacy and other goals of responsible AI development; for example, auditing models for potential unfairness typically requires sharing data externally (Field et al., 2023).

An alternative approach is laboriously creating anonymized data sets (e.g., Johnson et al. (2016a)). However, text data is extremely difficult to fully anonymize, and even lower dimensional data is often possible to re-identify with just small amounts of auxiliary data (Narayanan and Shmatikov, 2008; Sweeney, 2000). As anonymizing data typically involves masking sensitive information, this data is also not useful for tasks requiring sensitive information, such as developing a model to identify contact information for potential caretakers of a child (Field et al., 2023).

In this work, we consider *synthetic data* as an alternative approach, and we develop and evaluate methods for synthetic text generation. Our primary approach involves fine-tuning language models with differential privacy (DP), and using these models to generate synthetic text. DP offers a formal privacy guarantee and allows us to specify a privacy budget while preserving the utility of models trained on such data. Although the bulk of work in developing DP approaches has been centered around models trained on tabular and image-related data, there has been increasing interest in adapting the notion of DP to apply it to unstructured text data (Shi et al., 2022; Yue et al., 2021; Feyisetan et al., 2019).

A small amount of very recent work has similarly explored synthetic text for improving privacy

(Yue et al., 2023; Kurakin et al., 2023; Mattern et al., 2022a; Putta et al., 2023), but this work has lacked grounding in realistic applications, for example, running experiments with public internet data that language models may already have been exposed to during pre-training. In contrast, we conduct experiments on text data from two high stakes domains: healthcare and child protective services, and we rigorously evaluate the synthesized text for its utility, privacy, and potential fairness implications. For utility and privacy, we introduce novel well-motivated evaluation criteria ("silver" coreference modeling and entity-centric metrics). To the best of our knowledge, no prior work has investigated fairness considerations in this domain.

We do identify some promising opportunities for synthetic text, and we further these opportunities by proposing an NLI-based data-filtering approach to improve text coherence. However, our evaluations expose decreases in utility, potential privacy leakage, and potential unfairness, which are not well-reported in prior work.

Our primary contributions include: (1) a rigorous and reproducible evaluation framework that exposes limitations underestimated in prior work, (2) empirical results over real high stakes data, and (3) a proposed NLI-based data filtering approach to alleviate some of the limitations our evaluations expose. Overall, our work demonstrates that contrived metrics do not necessarily translate to more realistic scenarios, emphasizing the need for thorough in-domain evaluation to understand potential strengths and limitations of synthetic data.

2 Related Work

The majority of research on enabling shareable sensitive data has focused on text anonymization, replacing or redacting private information like names and addresses from text. While some approaches redact and replace sensitive information using deterministic rule-based systems (Mamede et al., 2016; Yermilov et al., 2023; Ben Cheikh Larbi et al., 2023; Sotolář et al., 2021; Volodina et al., 2020), others employ masked language models (Yermilov et al., 2023). Differentially private mechanisms have also been integrated into text sanitization processes, such as differentially private perturbation of text embeddings (Feyisetan et al., 2020) or sampling of replacement tokens (Yue et al., 2021; Chen et al., 2023) building on the principle of Metric-Local DP (Alvim et al., 2018). Although these methods are computationally inexpensive and domain-agnostic, they have weak privacy guarantees and limited capacity to modify text (Mattern et al., 2022b; Domingo-Ferrer et al., 2021; Brown et al., 2022).

Recently, datasets comprised entirely of synthetic data have become potentially viable (Guan et al., 2018; Yale et al., 2020). Our work differs from similar approaches to synthetic data generation in its focus on actual high stakes data and thorough grounded evaluation (Yue et al., 2023; Kurakin et al., 2023; Mattern et al., 2022a; Putta et al., 2023). Notably, Al Aziz et al. (2021) do similarly investigate healthcare data, but they do not evaluate potential privacy leakage, and their utility measures do not adequately capture errors in text fluency and consistency which is crucial for finer-grained applications.

A separate but overlapping line of work has focused on improving privacy in NLP models, rather than in generated data. This work has similarly trained NLP models with differential privacy but has evaluated direct performance of these models on downstream tasks (Li et al., 2021; Wu et al., 2022). Kurakin et al. (2023); Mattern et al. (2022a); Putta et al. (2023) connect these lines of work by offering training on DP-generated data as an alternative to DP-training on real data. Nevertheless, this line of work is not directly comparable to ours, given its differing goals.

3 Methodology

3.1 Text Generation

Our goal is to generate realistic, but entirely synthesized text for a high stakes domain, such as doctors' notes from a healthcare institution. We assume we have a data set of real text from that domain, which we can use to guide the generation.

Fine-tuning Our primary approach is to start with a pre-trained autoregressive language model (Xia et al., 2024), fine-tune it using the real in-domain data, and then generate new data from it. We utilize top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2019) to generate diverse synthetic notes, and we train a differentially private version of the model using DP-SGD. For reference, we provide background on DP and DP-SGD in Appendix A.

We condition the text generation on *control codes* (Keskar et al., 2019). More specifically, 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:

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

Controllable generation approaches enable the generation of notes with specific properties. We primarily use them to enable classification-based utility evaluations (described in §3.3).

ICL In order to explore the potential capabilities of much larger models and investigate if fine-tuning is actually needed, we also generate notes using in-context learning. We provide as context examples of training data text with pre-pended control codes, followed by an additional set of codes to prompt the model to generate content in accordance with the final set of codes. The number of examples provided varies, as we require that each control code for the note to be generated has at least one corresponding note within the examples that contains it. This approach is most directly comparable to the fine-tuned models without DP.

3.2 NLI-based data filtering

Synthetic text often contains inconsistencies and contradictions (see for example, samples provided in the appendix of Yue et al. (2023)). In order to improve the quality of generated text, we experiment with using natural language inference (NLI) models to score and filter out inconsistent text. NLI-based approaches have previously been used to rank or evaluate the quality of the generated text (Dušek and Kasner, 2020; Garneau and Lamontagne, 2021; Chen and Eger, 2023) and have been incorporated into the generation pipeline to enhance the consistency of outputs produced by LMs (Mersinias and Mahowald, 2023). Specifically, we use a pre-trained model fine-tuned over MNLI to predict entailment, neutral, and contradiction for each pair of consecutive sentences in the text. We then take the percentage of entailments and neutrals (e.g. the absence of contradictions) as the text's NLI score (denoted by S_{NLI}), which we use to rank and filter model outputs.

3.3 Utility Evaluation

Given our goal of developing synthetic data that could be shared externally with researchers or third-party contractors, we evaluate the data's utility based on the performance of NLP models trained over this data.

Classification Similar to prior work (Yue et al., 2023; Kurakin et al., 2023), we evaluate model performance over classification tasks, where we use the control codes provided during text generation as class labels. We focus on multiclass and/or multilabel classification tasks, and we compare model performance as task difficulty increases.

Coreference Resolution Classification tasks can be highly dependent on keywords and phrases, and they do not necessarily require training data to be coherent and consistent across a full paragraph or document. Consistency of entity properties across a document, however, is a necessary condition for coreference training data. Coreference and the related task of mention detection also offer a realistic use case in processing expert-written notes (Gandhi et al., 2023). Thus, we measure the utility of the synthetic data for training coreference models.

Unlike classification labels, coreference annotations cannot be easily generated through control codes. In a practical setting, annotations of coreference clusters would likely be conducted over synthesized data manually by hired annotators or researchers, but this process does not scale for evaluation of multiple iterations of synthetic data evaluation. Instead, we use a fine-tuned coreference model to simulate "silver" annotations over the synthesized data.

More specifically, given a subset of the original dataset D annotated with gold coreference clusters, we first finetune a pretrained coreference model (Kirstain et al., 2021) on this data. Using this model, we infer coreference clusters over synthetic data from the same domain which we consider silver annotations. We finetune a separate coreference model that has not been task-finetuned with the silver coreference clusters to approximate the utility of the synthetic data for coreference resolution.

We run all experiments with a neural coreference model (Kirstain et al., 2021). We report results after finetuning the model for 40 epochs, where scores are averaged over standard coreference metrics: MUC, B^3, CEAF$_{\phi_4}$.

3.4 Privacy Evaluation

Canary Attacks Consistent with prior work, to assess the potential leakage of sensitive information in our training data and the extent to which the model memorizes personally identifiable information (PII), we use the canary evaluation method proposed by (Carlini et al., 2018). This approach involves injecting artificial canary sequences containing PII into the training data and analyzing the likelihood of the frequency of appearance of this PII in the generated outputs.

We create artificial canary samples that are contextually relevant to both domains and include PII such as names, emails, addresses, and numeric identifiers (details in the appendix in Table 14 and Table 13). Following the methodology outlined in (Yue et al., 2023), we vary the number of injections of these canary samples into our training data for 1, 10, and 100 repetitions. For each canary, we generate 10,000 candidate sequences and rank the canaries based on their perplexity score.

Entity-focused metrics As canary evaluations are only a proxy for assessing potential privacy risks and may not be comprehensive, we directly leverage entity markers in our datasets to evaluate privacy concerns (we provide details on data-specific entity definitions in §4).

We compare the frequency of identified entities in the original vs. synthetic data. Further, while an isolated entity poses some privacy risk, the risk is magnified if the context surrounding the entity is also leaked. Thus we examine the frequency of entities with variable-length surrounding context in the synthetic data and compare them with the training data to estimate the number of memorized patterns that reappear in the synthetic data.

3.5 Fairness Evaluation

We compute fairness metrics over the same control-code classification tasks as the utility evaluation (§3.3). In data with available demographic information, we compare fairness classification for race and gender subgroups using equality difference (ED) and equalized odds (EO) metrics. For ED, for instance, False Positive Equality Difference (FPED) is the sum of the differences between the overall false positive rate (FPR) for the entire dataset and the FPR for each subgroup. EO constitutes a stricter notion of fairness by evaluating whether both the FPR and TPR rates are the same across all groups. In both cases, values closer to zero indicate that the model performs more uniformly across subgroups, with zero indicating perfect parity across subgroups. For reference, we formally define these metrics in Appendix C.

4 Experimental Setup

4.1 Data

Healthcare Our primary source of healthcare data is the MIMIC-III Clinical Database (Johnson et al., 2016b,a; Goldberger et al., 2000), which contains $>$ 2M deidentified notes associated with $>$ 40K patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

As control codes we use ICD-9 codes, which are a standardized format for medical conditions that have been human-annotated in MIMIC. Each note can contain multiple possible codes, making our evaluation task multiclass and multilabel. There are $>$ 8000 unique ICD-9 codes. Thus, we restrict data to notes containing any of the n most frequent ICD-9 codes, where we typically set $n = 10$ and report $n \in 3, 5$ for some comparisons, similar to Al Aziz et al. (2021); Huang et al. (2019). As a result, the training data size for the generative models can vary depending on the value of n. The dataset splits for the classification tasks are provided in Appendix D. To ensure synthetic data is balanced comparably to real data when evaluating fairness, we additionally provide the patient's ethnicity and biological sex as control codes.

For coreference resolution, we use notes from the MIMIC-II Database annotated for coreference as a part of the i2b2/VA Shared-Task and Workshop in 2011 (Uzuner et al., 2012). This includes 251 train documents, 51 of which we have randomly selected for development and 173 test documents.

As the MIMIC data is already deidentified, we directly leverage the strings used for deidentification, e.g. *[**Hospital1 18**]*, *[**First Name3 (LF) 2704**]*, in order to conduct entity-centric privacy evaluations. Finally, we note that although the MIMIC-III diagnoses notes are not permissible to be used for training publicly available language models, there remains a possibility that some MIMIC notes may have been indirectly included in the training data through various other sources.

Child Protective Services We additionally report results over a data set of contact notes from a county-level Department of Human Services (DHS). These contact notes log contact with families involved in child protective services, and they are written by caseworkers and other service providers. Unlike MIMIC-III, this data set is not deidentified, which makes it a more realistic test data set, but also prevents the data from being publicly accessible. Throughout our work, this data was stored on a secure server with restricted access, in accordance with IRB-approved protocol and a data sharing agreement established with the county.

The full data set contains 3.1M notes, from approximately 2010 to November 23, 2020. As control codes, we use existing metadata, specifically, the "Contact Source Description" field, which specifies one of five possible labels for each note. Similar to the ICD9 codes, we use the 3 most frequent labels: *Case*, *Investigation*, and *Call Screen*. For coreference resolution, we use a set of 200 notes annotated for coreference by prior work and shared with us by the county (Gandhi et al., 2023). This data has train/dev/test sets of sizes 100/10/90 notes. Finally, for entity-centric evaluations, we use a spaCy NER model to identify spans of entities in the text, and we focus on entities likely to contain private identifying information (e.g., names and organizations).

As CPS cases are complex and involve multiple people, the notion of race or gender for a note is less clear than in the MIMIC data. Thus, we do not report fairness results for this data. We also do not report ICL results, as our single secure server did not have sufficient resources for the larger model.

4.2 Models

Our primary text generation model is Sheared-LLaMA-1.3B[1]. We fine-tune using Low-Rank Adaption (LoRA) (Hu et al., 2021), and we use Opacus (Yousefpour et al., 2022) for DP fine-tuning. We generally set a privacy budget of $\epsilon = 8$, and δ = 1e-5 (considering our relatively small dataset size), and we report some results with $\epsilon = 4$ for comparison. For ICL, we used the instruction-tuned BioMistral-7B[2] model. As the inference for the BioMistral 7B model is expensive, we have generated a limited number of notes with which we carry out these experiments. For the NLI-based filtering, we use a pretrained BERT-base model fine-tuned over MNLI.[3] As this filtering is intended to improve text coherence, which is less important for classification, we report results from this approach using coreference metrics. For the classification tasks, we fine-tune a pretrained BERT-base model. We have specified the hyperarameters for each of the models used, dataset distributions and additional detail regarding the experimental setup in Appendix B and Appendix D.

[1]https://huggingface.co/princeton-nlp/Sheared-LLaMA-1.3B

[2]https://huggingface.co/BioMistral/BioMistral-7B

[3]https://huggingface.co/JeremiahZ/bert-base-uncased-mnli

5 Results

5.1 Utility

Overall Classification Tables 1 and 2 report results for classification tasks for all models, for the healthcare and CPS data respectively. Unsurprisingly, models trained on data generated from DP fine-tuned models generally under-perform models trained on real data or data generated without DP. Table 1 reports performance for varying task complexity by increasing number of labels n for our multilabel ICD-9 code classification task. For simpler tasks, e.g. ICD-$9_{n=3}$, there is a much smaller performance degradation and the $D_{\epsilon=\infty}$ (F1=0.87) and $D_{\epsilon=8}$ (F1=0.84) models are nearly comparable. In contrast, there is much larger performance degration for the more difficult ICD-$9_{n=10}$ task, where F1=0.61 for $D_{\epsilon=\infty}$ and F1=0.37 for $D_{\epsilon=8}$.

In the classification task with the CPS data (Table 2), however, we notice a significant drop in performance for models trained over $D_{\epsilon=8,\ 4}$. From examining the data, this task is generally more difficult and the associations between the administrative label and the text in the real data can be quite subtle. It is likely that the generative model often fails to pick up on these associations, and noise introduced by DP further masks these subtleties.

Overall Coreference Table 3 reports coreference results. For comparison, we report $D_{real(gold)}$, model performance when trained over gold in-domain data, which represents the best possible performance we can obtain with human annotations and $D_{real(sivler)}$, model performance when trained over silver annotated real data. The 15 point performance difference in F1 between these two setups represents the performance degradation we should expect to see as a result of inevitable cascading errors from the silver annotations.

Data generated without DP seems to outperform data generated with DP in both models, but the margin is larger for synthetic healthcare data. Since the mention detection performance for CPS data is much higher than for healthcare data, it is likely

Training Data	Dataset	F1 Micro	F1 Macro	Subset Accuracy
D_{real}	ICD-9$_{n=10}$	0.70 ± 0.010	0.67 ± 0.012	0.32 ± 0.016
$D_{\epsilon=\infty}$	ICD-9$_{n=10}$	0.66 ± 0.001 ↓-0.04	0.61 ± 0.003 ↓-0.06	0.26 ± 0.004 ↓-0.06
$D_{\epsilon=8}$	ICD-9$_{n=10}$	0.51 ± 0.007 ↓-0.19	0.37 ± 0.017 ↓-0.30	0.14 ± 0.007 ↓-0.18
D_{ICL}	ICD-9$_{n=10}$	0.57 ± 0.011 ↓-0.13	0.47 ± 0.014 ↓-0.20	0.21 ± 0.008 ↓-0.11
D_{real}	ICD-9$_{n=5}$	0.77 ± 0.008	0.76 ± 0.016	0.56 ± 0.007
$D_{\epsilon=\infty}$	ICD-9$_{n=5}$	0.75 ± 0.003 ↓-0.02	0.73 ± 0.003 ↓-0.03	0.55 ± 0.004 ↓-0.01
$D_{\epsilon=8}$	ICD-9$_{n=5}$	0.66 ± 0.006 ↓-0.11	0.57 ± 0.008 ↓-0.19	0.44 ± 0.007 ↓-0.12
D_{real}	ICD-9$_{n=3}$	0.89 ± 0.000	0.90 ± 0.000	0.76 ± 0.002
$D_{\epsilon=\infty}$	ICD-9$_{n=3}$	0.87 ± 0.001 ↓-0.02	0.87 ± 0.001 ↓-0.03	0.73 ± 0.006 ↓-0.03
$D_{\epsilon=8}$	ICD-9$_{n=3}$	0.84 ± 0.006 ↓-0.05	0.84 ± 0.007 ↓-0.06	0.68 ± 0.005 ↓-0.07

Table 1: Difference in performance between models trained on the synthetic data generated with ($D_{\epsilon=8}$) and without ($D_{\epsilon=\infty}$) DP and the models trained on real data (D_{real}) for multilabel ICD code classification with the top 10, 5, and 3 most frequent labels. Performance degradation greatly increases for more complex tasks.

	F-1 Score	Accuracy
D_{real}	0.78 ± 0.018	0.89 ± 0.006
$D_{\epsilon=\infty}$	0.64 ± 0.139 ↓0.14	0.86 ± 0.023 ↓0.03
$D_{\epsilon=8}$	0.29 ± 0.000 ↓0.49	0.78 ± 0.000 ↓0.11
$D_{\epsilon=4}$	0.32 ± 0.054 ↓0.46	0.79 ± 0.006 ↓0.10

Table 2: Difference in performance between models trained on data generated with differential privacy and models trained on real data, evaluated over CPS classification, for varying privacy budgets.

that the coreference score is dominated by the mention detection task, and consequently coreference performance is a weaker signal of coherence and consistency of entities for the CPS domain than the healthcare domain.

We inspect how the NLI-reranking approach to improve consistency is captured by the coreference model utility by comparing subsets of synthetic data ranked low and high for consistency. The reranking approach is consistent with coreference utility in that stronger coreference performance is associated with higher ranked examples. This result, however, does not hold for the CPS domain, likely as a result of the dominance of the mention detection task which makes consistency less relevant for strong coreference performance.

5.2 Privacy

Canary Attacks Table 4 reports results for canary attacks. The DP fine-tuned models exhibit higher perplexity scores for all the canaries, demonstrating that models trained with DP are less likely to output phrases from training data. DP similarly improves (increases) rank for most canaries, again indicating that models trained with DP are less likely to output phrases from training data. Although Yue et al. (2023) assert that the differentially private training of language models can effectively *eliminate* the risk of privacy leakage, our canary evaluation results indicate that this may not hold true for all types of PII. This is further illustrated by our experiments over PII in our entity-centric analysis.

Entity-centric Metrics We do not assess the leakage (e.g., appearance in generated text) of the canaries in our generated sequences in this work. Instead, we perform this analysis over leakage of the PII that is already present in the training data.

The entity-centric metrics (Table 5) show that while DP-generated data does contain fewer instances of potentially sensitive information, these entities are not removed from the data entirely, and there is still the risk of leakage. In Table 6, we gauge how often sequences containing these leaked entities appear in the generated outputs, where we vary the number of words in the context surrounding the entities (denoted by $k \in \{1, 2, 3, 4\}$). The results provide further evidence that, while training models with differential privacy may decrease the risk of information memorization, it does not provide a failsafe. There is a notable disparity in

Training Data	Healthcare		CPS	
	Mention Detection	**Coreference**	**Mention Detection**	**Coreference**
$D_{real(gold)}$	0.799 ± 0.013	0.703 ± 0.011	0.877 ± 0.004	0.789 ± .005
$D_{real(silver)}$	0.659 ± 0.121	0.552 ± 0.126	0.805 ± 0.007	0.642 ± 0.008
$D_{\epsilon=\infty}$	0.596 ± 0.014	0.422 ± 0.014	0.785 ± 0.001	0.594 ± 0.014
$D_{\epsilon=\infty}^{S_{NLI},\ high}$	0.588 ± 0.069	0.430 ± 0.072	0.756 ± 0.003	0.571 ± 0.005
$D_{\epsilon=\infty}^{S_{NLI},\ low}$	0.462 ± 0.034	0.288 ± 0.027	0.750 ± 0.046	0.566 ± 0.046
D_{ICL}	0.712 ± 0.010	0.588 ± 0.022	-	-
$D_{\epsilon=8}$	0.575 ± 0.002	0.404 ± 0.023	0.777 ± 0.013	0.582 ± 0.020
$D_{\epsilon=8}^{S_{NLI},\ high}$	0.570 ± 0.086	0.423 ± 0.087	0.783 ± 0.002	0.593 ± 0.008
$D_{\epsilon=8}^{S_{NLI},\ low}$	0.496 ± 0.083	0.335 ± 0.063	0.785 ± 0.006	0.598 ± 0.008

Table 3: F1 scores for coreference and mention detection over entities from human-annotated test splits of the CPS and i2b2/VA datasets. All synthetic datasets are annotated with silver labels. We compare performance between synthetic data generated from models where ($\epsilon = 8,\ \infty$) and models trained with real data. We also compare the performance over data generated from these models with a high S_{NLI} score and a low S_{NLI} score.

		Rank	Perplexity
Healthcare	Name	10001 / 10001	54.06 / 50.11
	Address	5645 / 3088	62.57 / 41.08
	Number	1 / 1	14.59 / 9.54
	Email	9479 / 9372	71.98 / 37.40
CPS	Name	1 / 1	12.355 / 12.142
	Address	9863 / 7849	26.741 / 21.726
	Number	9999 / 9645	26.038 / 16.409
	Email	10000 / 9951	87.724 / 52.070

Table 4: Rank and perplexity metrics for 10-insertion canary attacks over MIMIC and CPS data (1 and 100 insertions, reported in Appendix E, are similar). Each column is formatted as $\epsilon = 8/\epsilon = \infty$. DP reduces but does not eliminate privacy risks for all canaries.

the frequency of phrases from the training data reproduced in these datasets: $D_{\epsilon=\infty}$ contains nearly 2.6 times as many phrases as the $D_{\epsilon=8}$, but the phrase leakage from $D_{\epsilon=8}$ is still non-zero. On the other hand, while D_{ICL} is 0.6 times the size of the other datasets, it seems to regurgitate contextual information about these entities from the in-context samples less frequently. However, results from Table 5 indicate that it still poses privacy risks, as the ICL tends to reproduce these entities, even if not the contexts in which they appear.

5.3 Fairness

We report the FNED and Equalized Odds (EO) metrics for the results from the ICD-9$_{n=10}$ multilabel classification tasks in Table 8. The metrics reflect the difference in model performance for the gender and race/ethnicity subgroups with more than 100 samples in the test set, with a larger value indicating more disparate performance across the subgroups. While the gender metrics indicate minimal performance differences, the race/ethnicity metrics show significant disparities. The disparate performance increases for models trained over the data generated from the DP model ($D_{\epsilon=8}$) as compared to the model without DP ($D_{\epsilon=\infty}$). Although D_{ICL} appears to preserve utility for coreference resolution and mention detection (Table 3), and provides better utility than $D_{\epsilon=8}$ for classification, it is consistently exhibits the most disparate performance over subgroups. We report additional fairness metrics in Appendix C in Table 8 that show similar trends.

6 Discussion

Overall, our results are consistent with prior work in that we find only small performance degradation when training a model on DP-generated synthetic text as compared to real data for relatively less fine-grained (e.g. ICD-9$_{n=3}$, in Table 1) classification tasks. Similarly, we do find evidence that DP reduces potential privacy leakage in that artificial canaries (Table 4) and real entities (Table 5) are generated less frequently by DP-fine-tuned models.

However, our evaluations also expose previously unexplored weaknesses to this approach. For instance, the model performance's generally de-

	Healthcare							CPS		
	Overall	**Name**	**Loc.**	**Hospital**	**DT**	**NI**	**OI**	**Org.**	**Person**	**Date**
D_{real}	1617.17	797.51	64.71	109.13	53.56	318.94	273.32	968.90	1419.12	194.64
$D_{\epsilon=\infty}$	111.06	88.62	3.00	11.99	0.93	2.48	4.03	88.93	62.57	29.36
D_{ICL}	96.18	60.85	7.74	12.22	1.27	5.04	9.06	-	-	-
$D_{\epsilon=8}$	48.42	40.04	1.13	7.22	0.03	0.00	0.00	25.99	12.89	34.17

Table 5: Entity-centric privacy evaluation. We report the number of instances of each type of identifier in the real or generated data, divided by the total number of notes, multiplied by 1000. Results can be read as the number of identifiers estimated to occur in 1000 notes of this type. "DT" stands for Date/Time, "NT" refers to numeric identifiers, such as phone number, social security number, etc., and "OI" reports other identifiers.

	Healthcare		CPS	
	Ratio	**Count**	**Ratio**	**Count**
$D_{\epsilon=\infty}$	0.00504	16271	0.01854	5150
D_{ICL}	0.00117	3761	-	-
$D_{\epsilon=8}$	0.00196	6316	0.00416	1010
$D_{\epsilon=4}$	-	-	0.00436	1069

Table 6: Unique contexts in which entities (PER/ORG categories for CPS) in the real data appear in the synthetic data. Surrounding context word lengths vary from 1 to 4. "Count" denotes the number of entity-contexts appearing in both the generated data and real data. "Ratio" denotes that count divided by the number of phrases in either data.

		FNED	**Equalized Odds**
Race	D_{real}	0.34 ± 0.011	0.21 ± 0.009
	$D_{\epsilon=\infty}$	0.37 ± 0.006	0.21 ± 0.002
	D_{ICL}	**0.52 ± 0.008**	**0.30 ± 0.002**
	$D_{\epsilon=8}$	0.48 ± 0.034	0.28 ± 0.014
Gender	D_{real}	0.04 ± 0.005	0.04 ± 0.005
	$D_{\epsilon=\infty}$	0.03 ± 0.006	0.03 ± 0.006
	D_{ICL}	0.04 ± 0.005	0.04 ± 0.005
	$D_{\epsilon=8}$	0.03 ± 0.003	0.03 ± 0.003

Table 7: Fairness evaluation for the MIMIC-III ICD-$9_{n=10}$ task, for the gender and race categories.

grades much more sharply as task complexity increases (e.g. ICD-$9_{n=10}$ classification, Table 1), and there is still a substantial risk of data leakage (Tables 4-6). These results suggest that DP-generated synthetic data may be of sufficient quality for certain NLP tasks and domains, but the quality degradation from DP can be a limitation. Our NLI-based ranking suggests that some output text is higher quality than others, and further filter methods may offer opportunities to improve quality.

Further, simply applying DP during fine-tuning is not sufficient to prevent data leakage and more care needs to be taken. It may be possible to alleviate privacy risk through modifications to the pipeline, such as using stricter privacy budgets. A more promising approach may be to combine privacy-preserving techniques.

We further find substantial variance not only in the task difficulty, but also across data sets. While models perform comparably when trained on silver coreference annotations over synthetic text and real text for the CPS data, the synthetic data is markedly worse the than real data for MIMIC (Table 3). These differences could be due to a number of factors, such as the similarity between each private data set and the model pre-training data. Regardless, these results emphasize the importance of evaluating on in-domain data, as results are not likely to generalize.

Conclusions Our findings show that while DP reduces privacy risks, it does not eliminate them. The utility of synthetic data may not be comparable to real data for more complex tasks and may even introduce fairness issues. We also demonstrate that maintaining the coherence & consistency of synthetic text can benefit tasks like coreference resolution. While DP shows promise in these applications, our findings also indicate that additional elements need to be incorporated in the pipeline to potentially improve the quality and privacy-preserving aspects of synthetic data.

7 Limitations

The primary limitation of our work is the impossibility of considering all possible model and parameter configurations. While we selected highperforming models that we were able to fine tune and evaluate on our compute resources, results may differ for different pre-trained language models. Similarly, while we select hyper-parameters based on prior work and conduct some ablation studies, text-generation is extremely compute-intensive and a fully exhaustive hyper-parameter sweep is not feasible. Overall our results emphasize the need to thoroughly evaluate models on target data and cannot necessarily be assumed to generalize to untested data.

There are also additional approaches we do not explore that could reduce privacy risk or improve the quality of synthetic data generated during training. Examples include combining text-anonymization with DP fine-tuning or selective constraints applied to the training data to reduce the frequency of entity mentions. However, this is difficult in practice, as real-world data is complex with, for example, the same people mentioned across multiple CPS cases.

8 Ethical Considerations

Our work involves the use of private sensitive data, particularly the CPS data, which is not de-identified. To minimize risk, throughout this project we maintained a high level of data security, in compliance with IRB-approved protocol. The CPS data was exclusively stored on a secure restricted-access server with HIPPA-standard of security. All CPS experiments were conduct on this server, which also limited the models we could investigate. Our paper does not include any examples from either data set, in compliance with their respective data use agreements.

References

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

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 Trans. Comput. Healthcare*, 3(1).

Mário Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Invited paper: Local differential privacy on metric spaces: Optimizing the trade-off with utility. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 262–267.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2023. Clinical text anonymization, its influence on downstream NLP tasks and the risk of re-identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111, Dubrovnik, Croatia. Association for Computational Linguistics.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. 2018. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *CoRR*, abs/1802.08232.

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.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM*, 64(7):33–35.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

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. *Preprint*, arXiv:1805.04833.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2019. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. *CoRR*, abs/1910.08902.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 178–186, New York, NY, USA. Association for Computing Machinery.

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*, pages 1479–1492.

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.

Nicolas Garneau and Luc Lamontagne. 2021. Trainable ranking models to evaluate the semantic accuracy of data-to-text neural generator. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. 2021. Numerical composition of differential privacy. *Preprint*, arXiv:2106.02848.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. 2019. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*, 177:141–153.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2.

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. 2016b. 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. *Preprint*, arXiv:1909.05858.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Nuno Mamede, Jorge Baptista, and Francisco Dias. 2016. Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022a. 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.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022b. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Michail Mersinias and Kyle Mahowald. 2023. For generated text, is NLI-neutral text the best text? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2596–2602, Singapore. Association for Computational Linguistics.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125.

Bhanudas Suresh Panchbhai and Varsha Makarand Pathak. 2022. A systematic review of natural language processing in healthcare. *Journal of Algebraic Statistics*, 13(1):682–707.

Pranav Putta, Ander Steele, and Joseph W Ferrara. 2023. Differentially private conditional text generation for synthetic data production.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Ondřej Sotolář, Jaromír Plhák, and David Šmahel. 2021. Towards personal data anonymization for social messaging. In *Text, Speech, and Dialogue*, pages 281–292, Cham. Springer International Publishing.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health*, 671.

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.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinwei Wu, Li Gong, and Deyi Xiong. 2022. Adaptive differential privacy for language model training. In *Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022)*, pages 21–26, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared llama: Accelerating language model pre-training via structured pruning. *Preprint*, arXiv:2310.06694.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. 2022. Opacus: User-friendly differential privacy library in pytorch. *Preprint*, arXiv:2109.12298.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

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.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Background: Differential Privacy

Differential privacy 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, differential privacy 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 $F : D \rightarrow Y$ 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. If 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 or Laplacian 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 Hyperparameters

For training the autoregressive model, we used a batch size of 4, set the maximum sequence length to 1024 tokens and a batch size of 4. Training was conducted over 3 epochs with a learning rate set to 3e-4, optimized using the AdamW optimizer using the default hyperparameters for the same. For the LoRA hyperparameters, we used a dimension of 4 and an alpha value of 32, specifically targeting the query (q_proj) and value (v_proj) projection layers of the transformer. To ensure training stability, we applied gradient clipping with a maximum gradient norm of 1.0. For the DP fine-tuning of the autoregressive model, we train with a privacy budget of epsilon = 8, and considering our relatively small dataset size we set delta to 1e-5 for our experiments.

For training the downstream classifier, we conducted training over 3 epochs with a batch size of 8 and a maximum sequence length of 512 tokens. We utilized the AdamW optimizer with a learning rate of 5e-5.

During inference, we set the top-k sampling parameter to k = 50 and the nucleus sampling parameter to p = 0.95. We generate approximately 30k and 31k samples for the child welfare data and diagnosis notes for the 10 most frequent ICD-9 codes, respectively, which are then used to train the downstream classifiers. We use similar inference hyperparameters for the instruction-tuned BioMistral-7B model for ICL, we set the top-k value to 50, top-p to 0.9 and the penalty-alpha parameter to 0.6.

Our experiments for all the aforementioned experimental setups used an A100 GPU for the MIMIC data and A6000 GPUs on a single secure server for the CPS data.

C Fairness

The False Positive Equality Difference (FPED) metric is the sum of the differences between the overall false positive rate (FPR) for the entire dataset and the FPR for each subgroup $d \in D$, where D is a set consisting of all subgroups corresponding to a demographic attribute within the dataset.

$$\text{FPED} = \sum_{d=1}^{D} |\text{FPR}_{\text{overall}} - \text{FPR}_d| \tag{2}$$

$$\text{TNED} = \sum_{d=1}^{D} |\text{TNR}_{\text{overall}} - \text{TNR}_d| \tag{3}$$

Similarly, these ED metrics can be estimated for the true positive, true negative and false negative rates to estimate the TPED, TNED and FNED respectively. Lower values of these ED scores indicate that the model's performance is more consistent across different subgroups.

The Equalized Odds ratio is calculated as follows:

	FNED	FPED	TPED	TNED	Equalized Odds
Race					
D_{real}^{base}	0.34 ± 0.011	0.01 ± 0.003	0.34 ± 0.011	0.01 ± 0.003	0.21 ± 0.009
$D_{\epsilon=\infty}$	0.37 ± 0.006	0.01 ± 0.002	0.37 ± 0.006	0.01 ± 0.002	0.21 ± 0.002
$D_{\epsilon=8}$	0.48 ± 0.034	0.02 ± 0.007	0.48 ± 0.034	0.02 ± 0.007	0.28 ± 0.014
D_{ICL}	**0.52 ± 0.008**	**0.03 ± 0.003**	**0.52 ± 0.008**	**0.03 ± 0.003**	**0.30 ± 0.002**
Gender					
D_{real}^{base}	0.04 ± 0.005	0.00 ± 0.001	0.04 ± 0.005	0.00 ± 0.001	0.04 ± 0.005
$D_{\epsilon=\infty}$	0.03 ± 0.006	0.00 ± 0.001	0.03 ± 0.006	0.00 ± 0.001	0.03 ± 0.006
$D_{\epsilon=8}$	0.03 ± 0.003	0.00 ± 0.001	0.03 ± 0.003	0.00 ± 0.001	0.03 ± 0.003
D_{ICL}	0.04 ± 0.005	0.00 ± 0.001	0.04 ± 0.005	0.00 ± 0.001	0.04 ± 0.005

Table 8: Fairness evaluation for the MIMIC-III ICD-9$_{n=10}$ task, for the gender and race categories.

$$\mathrm{EO}_D = \max\Bigg(\max_{i\in D}(\mathrm{TPR}_i) - \min_{i\in D}(\mathrm{TPR}_i), \\ \max_{i\in D}(\mathrm{FPR}_i) - \min_{i\in D}(\mathrm{FPR}_i)\Bigg)$$

We have two categories of subgroups that are present in the MIMIC-III dataset over which we perform fairness evaluations with the downstream classifier trained over synthetic data with demographic control codes. The following categorical variables assigned to each within the dataset:

- **Gender:** Female, Male
- **Race/Ethnicity:** American Indian/Alaska Native, Asian, Black, Hispanic/Latino, Middle Eastern, Multi Race/Ethnicity, Other, Portuguese, South American, White

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.

D Data Statistics

Our train/dev splits for the CPS, ICD-9$_{n=10}$, ICD-9$_{n=5}$ and ICD-9$_{n=3}$ datasets the generative model was trained on are 90250/4750, 44215/2327, 37245/1960, 31317/1648 respectively.

The train/dev sets for the models trained for downstream classification on the real (D_{real}) and synthetic ($D_{\epsilon=\infty,\ 8,\ 4}$) CPS data are 18000/4875 and 27000/3000 respectively. The size of the test set for this task was 4875.

For the ICD-9$_{n=10}$ multilabelling task, the real (D_{real}) and synthetic ($D_{\epsilon=\infty,\ 8,\ 4}$) train/dev split was the same, with $\simeq$ 27920/3100 for all models, and the test set size was $\simeq$ 7500 samples. For the ICD-9$_{n=5}$ task, the train/dev split was the same for all models $\simeq$ 23520/2615, and the test set size was $\simeq$ 6315 samples. Similarly, for the ICD-9$_{n=3}$ task, the train/dev split was $\simeq$ 19780 / 2200, and the test set size was $\simeq$ 5310 samples. Each of these experiments for the downstream tasks (coreference/mention detection & classification) was averaged over 3 runs.

We report additional data statistics in Table 9 and Table 10.

E Extended Privacy Evaluation results

In Table 4 we report the full set of canary results (for 1, 10, and 100 insertions, for each canary type). Results are generally similar across different numbers of insertions, in that DP generally reduces rank and perplexity, thus improving privacy, but does not eliminate all risk of leakage.

Model	Mean TTR	1-gram Overlap Ratio	2-gram Overlap Ratio	3-gram Overlap Ratio	4-gram Overlap Ratio
$D^{base}_{(real,\text{ ICD-9}_{n=10})}$	0.474	0.827	0.805	0.773	0.750
$D_{(\epsilon=\infty,\text{ ICD-9}_{n=10})}$	0.569	0.165	0.083	0.046	0.028
$D_{(\epsilon=8,\text{ ICD-9}_{n=10})}$	0.539	0.153	0.085	0.043	0.022
$D_{(ICL,\text{ ICD-9}_{n=10})}$	0.448	0.134	0.093	0.051	0.028
$D^{base}_{(real,\text{ ICD-9}_{n=5})}$	0.468	0.679	0.639	0.585	0.548
$D_{(\epsilon=\infty,,\text{ ICD-9}_{n=5})}$	0.569	0.161	0.079	0.042	0.026
$D_{(\epsilon=8,\text{ ICD-9}_{n=5})}$	0.528	0.143	0.078	0.038	0.020
$D^{base}_{(real,\text{ ICD-9}_{n=3})}$	0.474	0.608	0.556	0.494	0.453
$D_{(\epsilon=\infty,\text{ ICD-9}_{n=3})}$	0.574	0.146	0.073	0.039	0.023
$D_{(\epsilon=8,\text{ ICD-9}_{n=3})}$	0.543	0.142	0.074	0.036	0.019

Table 9: Comparison of MIMIC-III TTR (Type-Token Ratio) and n-gram overlap statistics, with the overlap measured between unique n-grams in the synthetic data and the training data.

Data	Mean TTR	1-gram Overlap Ratio	2-gram Overlap Ratio	3-gram Overlap Ratio	4-gram Overlap Ratio
D^{base}_{real}	0.512	0.407	0.354	0.288	0.248
$D_{\epsilon=\infty}$	0.429	0.150	0.129	0.097	0.063
$D_{\epsilon=8}$	0.403	0.131	0.113	0.082	0.052
$D_{\epsilon=4}$	0.405	0.134	0.114	0.083	0.051

Table 10: Comparison of CPS TTR (Type-Token Ratio) and n-gram overlap statistics, with the overlap measured between unique n-grams in the synthetic data and the training data.

Model	Data Size	Phrase Overlap Ratio	Total # of Phrase Overlap	Total # of Phrases	Total # of Deidentified Phrases Generated
$D_{(real,\text{ ICD-9 }_{n=10})}$	44215	1	2935955	2935955	3845112
$D_{(\epsilon=\infty,\text{ ICD-9}_{n=10})}$	31020	0.00504	16271	3229278	369390
$D_{(\epsilon=8,\text{ ICD-9}_{n=10})}$	31020	0.00196	6316	3229096	322280
$D_{(ICL,\text{ ICD-9}_{n=10})}$	19640	0.00117	3761	3205098	316905
$D_{(real,\text{ ICD-9}_{n=5})}$	37245	1	2565699	2565699	3352588
$D_{(\epsilon=\infty,\text{ ICD-9}_{n=5})}$	26136	0.00478	13537	2831658	323963
$D_{(\epsilon=8,\text{ ICD-9}_{n=5})}$	26136	0.00160	4554	2841128	297966
$D_{(real,\text{ ICD-9}_{n=3})}$	31317	1	2066294	2066294	2697186
$D_{(\epsilon=\infty,\text{ ICD-9}_{n=3})}$	21978	0.00502	11438	2280229	267603
$D_{(\epsilon=8,\text{ ICD-9}_{n=3})}$	21978	0.00174	3942	2267090	217834

Table 11: Analysis for the MIMIC-III dataset of all the unique contexts in which entities of from all categories from the training data appear in the synthetic data, considering surrounding context word lengths varying from 1 to 4. D_{real} corresponds to the training data the generative models were trained on.

	MIMIC		CPS	
	Rank	**Perplexity**	**Rank**	**Perplexity**
100 Insertions				
Name	10001 / 10001	53.14 / 34.51	1 / 1	11.516 / 8.659
Address	5535 / 35	61.57 / 20.13	9907 / 1	33.909 / 4.117
Number	1 / 1	14.48 / 7.84	9999 / 4696	25.337 / 8.712
Email	9455 / 1	70.15 / 11.01	10000 / 140	90.389 / 14.626
10 Insertions				
Name	10001 / 10001	54.06 / 50.11	1 / 1	12.355 / 12.142
Address	5645 / 3088	62.57 / 41.08	9863 / 7849	26.741 / 21.726
Number	1 / 1	14.59 / 9.54	9999 / 9645	26.038 / 16.409
Email	9479 / 9372	71.98 / 37.40	10000 / 9951	87.724 / 52.070
1 Insertion				
Name	10001 / 10001	54.20 / 52.46	1 / 1	12.374 / 12.931
Address	5696 / 6590	62.82 / 57.51	9867 / 8770	26.808 / 25.666
Number	1 / 1	14.63 / 9.87	9999 / 9653	26.041 / 17.599
Email	9549 / 9745	72.29 / 43.92	10000 / 10000	87.987 / 75.408
0 Insertions				
Name	10001 / 10001	60.164 / 52.682	1 / 1	12.375 / 13.052
Address	6256 / 7298	57.133 / 65.527	9867 / 9080	26.814 / 26.142
Number	1 / 1	14.440 / 11.356	9999 / 9541	26.040 / 17.262
Email	9594 / 9694	66.829 / 41.371	10000 / 10000	88.038 / 81.019

Table 12: Rank and perplexity metrics for canary attacks over MIMIC and CPS data. Each column is formatted as $\epsilon = 8/\epsilon = \infty$. Perplexity scores suggest that DP reduces privacy metrics for all canaries, and generally show similar privacy improvements.

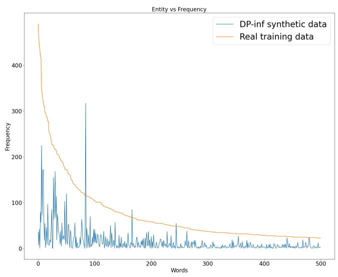

Figure 1: $D_{\epsilon=\infty}$

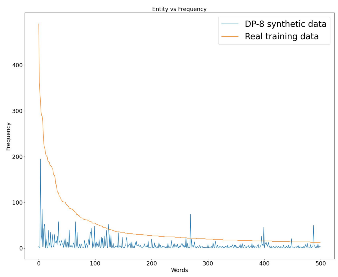

Figure 2: $D_{\epsilon=8}$

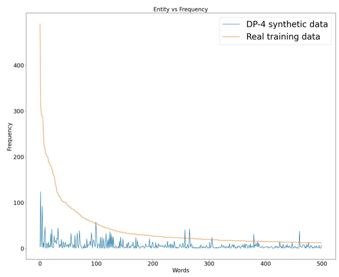

Figure 3: $D_{\epsilon=4}$

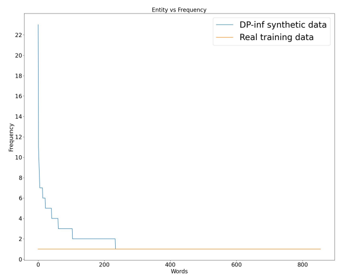

Figure 4: $D_{\epsilon=\infty}$

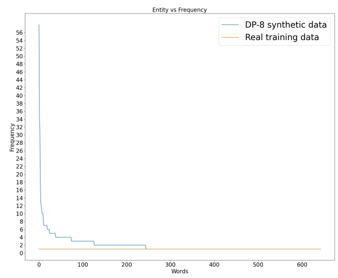

Figure 5: $D_{\epsilon=8}$

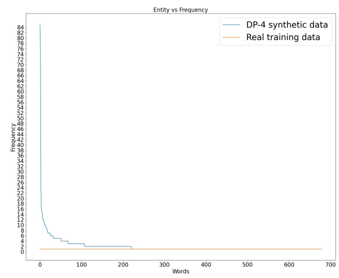

Figure 6: $D_{\epsilon=4}$

Figure 7: CPS data: Graph depicts the frequency of overlapping entities between the training data D_{train} for the generative model and synthetic data. The top row presents the top 500 most frequent entities from each dataset, limited to entities with a frequency count below 500 in D_{train}. The bottom row includes all instances of entities found exclusively in the synthetic data, where the count in D_{train} equals 1.

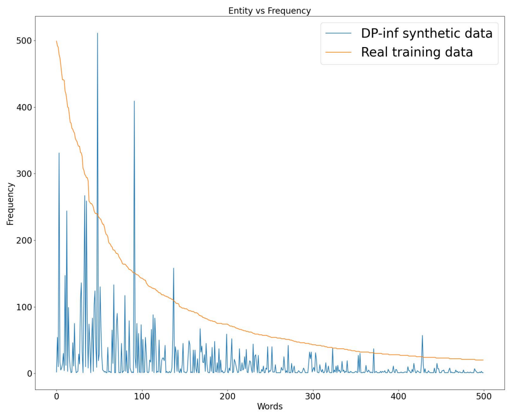

Figure 8: $D_{\epsilon=\infty}$

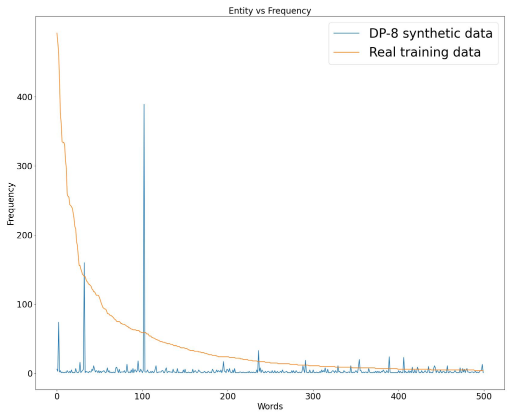

Figure 9: $D_{\epsilon=8}$

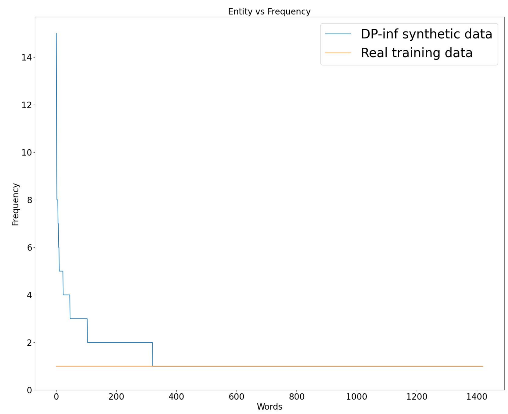

Figure 10: $D_{\epsilon=\infty}$

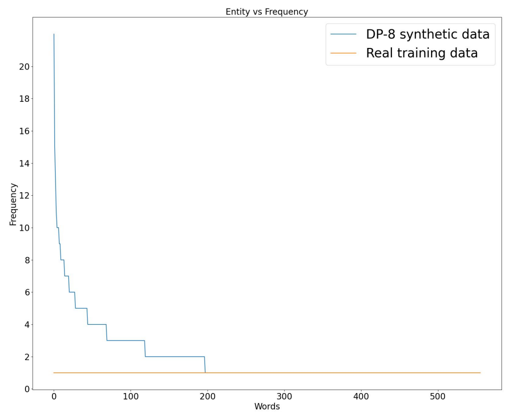

Figure 11: $D_{\epsilon=8}$

Figure 12: MIMIC-III ICD-9$_{n=10}$ data: Graph depicts the frequency of overlapping entities between the training data D_{train} for the generative model and synthetic data. The top row presents the top 500 most frequent entities from each dataset, limited to entities with a frequency count below 500 in D_{train}. The bottom row includes all instances of entities found exclusively in the synthetic data, where the count in D_{train} equals 1.

Type	Canary Sequence
Name	Pt Robin Kramer is has been admitted to the hospital and the patient's medical records were updated yesterday.
Address	The patient has been admitted to the hospital at 789 Elmford St, Maplegreen.
Number	The patient's contact number for emergencies is (541) 545-6789.
E-mail	In case of any inquiries, the patient can be contacted at ptcareastps456@email.com.

Table 13: The canaries inserted into the training data for the models fine-tuned to generate synthetic MIMIC-III data.

Type	Canary Sequence
Name	Robin Kramer has been placed in temporary foster care and there is an ongoing investigation into the child's welfare.
Address	The CW visited the foster family's address at 456 W Oak Avenue, Springfield, IL.
Number	The case number CW-2023-56893 has been assigned for tracking purposes.
E-mail	The CW can contact the foster family at randuser789@xyzreportnews.com in case of any emergencies.

Table 14: The canaries inserted into the training data for the models fine-tuned to generate synthetic CPS data.

Model	Data Size	Overlap Ratio of EP	Overlap EP #	Total EP # in D_{real} + $D_{synth-data}$	Total EP # in $D_{synth-data}$
D_{real}	90250	1.00000	216592	216592	326926
$D_{\epsilon=\infty}$	30000	0.01854	5150	277710	105213
$D_{\epsilon=8}$	30000	0.00416	1010	242528	34153
$D_{\epsilon=4}$	30000	0.00436	1069	244932	37787

Table 15: Analysis for the CPS data of all the unique contexts in which entities in the PERSON/ORG categories from the training data appear in the synthetic data, considering surrounding context word lengths varying from 1 to 4. D_{real} corresponds to the training data the generative models were trained on.

Model	Data Size	Overlap Ratio of EP	Overlap EP #	Total EP # in D_{real} + $D_{synth-data}$	Total EP # in $D_{synth-data}$
D_{real}	90250	1	323272	323272	471245
$D_{\epsilon=\infty}$	30000	0.01817	7410	407779	164060
$D_{\epsilon=8}$	30000	0.00678	2504	369061	64519
$D_{\epsilon=4}$	30000	0.00580	2129	366935	61565

Table 16: Analysis for the CPS data of all the unique contexts in which entities of from all categories from the training data appear in the synthetic data, considering surrounding context word lengths varying from 1 to 4. D_{real} corresponds to the training data the generative models were trained on.