Differentially Private Synthetic Data via Foundation Model APIs 2: Text

Chulin Xie [1] **Zinan Lin** [2] **Arturs Backurs** [2] **Sivakanth Gopi** [2] **Da Yu** [3] **Huseyin Inan** [2] **Harsha Nori** [2]
Haotian Jiang [2] **Huishuai Zhang** [2] **Yin Tat Lee** [2] **Bo Li** [1 4] **Sergey Yekhanin** [2]

Abstract

Text data has become extremely valuable due to the emergence of machine learning algorithms that learn from it. A lot of high-quality text data generated in the real world is private and therefore cannot be shared or used freely due to privacy concerns. Generating synthetic replicas of private text data with a formal privacy guarantee, i.e., differential privacy (DP), offers a promising and scalable solution. However, existing methods necessitate DP finetuning of large language models (LLMs) on private data to generate DP synthetic data. This approach is not viable for proprietary LLMs (e.g., GPT-3.5) and also demands considerable computational resources for open-source LLMs. Lin et al. (2024) recently introduced the *Private Evolution* (PE) algorithm to generate DP synthetic images with only API access to diffusion models. In this work, we propose an augmented PE algorithm, named AUG-PE, that applies to the complex setting of text. We use API access to an LLM and generate DP synthetic text without any model training. We conduct comprehensive experiments on three benchmark datasets. Our results demonstrate that AUG-PE produces DP synthetic text that yields competitive utility with the SOTA DP finetuning baselines. This underscores the feasibility of relying solely on API access of LLMs to produce high-quality DP synthetic texts, thereby facilitating more accessible routes to privacy-preserving LLM applications. Our code and data are available at https://github.com/AI-secure/aug-pe.

1. Introduction

With recent advances in natural language processing (NLP), text-based applications have greatly facilitated our lives. These include AI-assisted medical record summaries (Rumshisky et al., 2016), email and document autocomplete tools (Voytovich & Greenberg, 2022; CNN, 2023), and personalized chatbots (Chew, 2022). However, all these applications (among others) rely on collecting private text data from users to train LLMs, which raises serious privacy concerns as LLMs may memorize and leak sensitive information about users (Carlini et al., 2021; Lukas et al., 2023; Wang et al., 2023). Differentially private synthetic text is a promising and actively studied solution (Putta et al., 2022; Bommasani et al., 2019). It aims to create a new text dataset with similar characteristics to the original private data while ensuring privacy by protecting sensitive information in each sample (known as Differential Privacy (DP) (Dwork et al., 2014)). The DP synthetic text can then be used in developing any downstream NLP system without adding extra privacy risks. It also allows the safe sharing of private data more broadly. For example, hospitals can share their private medical data for research purposes by creating a DP synthetic version of their data.

The state-of-the-art DP synthetic text approach is to *finetune pretrained generative language models (LMs) on private data* with DP-SGD (a DP variant of SGD (Abadi et al., 2016)) (Yue et al., 2023; Kurakin et al., 2023; Mattern et al., 2022a) (short-handed as *DP finetune generator*; see Fig. 1). Unlike non-DP ML applications, which have been greatly advanced by powerful LLMs such as GPT-4 (OpenAI, 2023b) and LLaMA (Touvron et al., 2023a;b) in a short time after they are released, the state-of-the-art DP synthetic text approaches are unfortunately still based on GPT-2.[1] The reasons are: (1) Many powerful LLMs such as GPT-4, Claude, and Bard are only accessible through APIs. DP finetuning them is not feasible.[2] (2) Even though some LLMs (e.g., LLaMA) are open-source, finetuning them with

[1]University of Illinois Urbana-Champaign [2]Microsoft Research [3]Sun Yat-sen University [4]University of Chicago. Correspondence to: Chulin Xie <chulinx2@illinois.edu>, Zinan Lin <zinanlin@microsoft.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] The 175 billion-parameter GPT-3 has also been used for DP synthetic text (He et al., 2022). However, the solution is not publicly accessible as GPT-3 is proprietary.

[2] Although standard finetuning APIs are provided for some of the models (OpenAI, 2023a), DP finetuning requires a special implementation and no model provides this custom API to date.

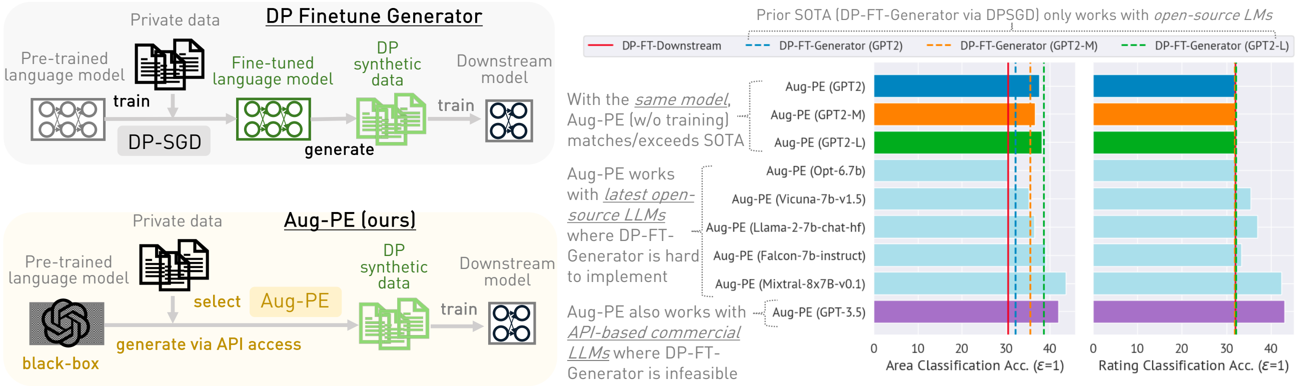

Figure 1: Instead of finetuning LLMs with DP-SGD to generate synthetic text, our AUG-PE only requires inference APIs of LLMs. AUG-PE works with the latest open-source LLMs and API-based LLMs to generate DP synthetic text with improved utility on OpenReview dataset, where DP-SGD finetuning is either hard to implement or infeasible.

DP is resource-intensive and non-trivial to implement due to the need to calculate *per-sample* gradients (see § 2).

A recent DP synthetic data framework called Private Evolution (PE) (Lin et al., 2024) offers a new opportunity to circumvent these challenges by only requiring API access to foundation models, without needing any model training. The high-level idea is to first draw random samples from a foundation model, and then iteratively improve them by selecting (with DP) the most similar ones to the private dataset and querying foundation models to generate more of such samples. PE shows promising results on *images* by leveraging pretrained Diffusion Models (Rombach et al., 2022): in certain cases, PE achieves an even better privacy-utility trade-off than DP finetuned generators (Lin et al., 2024).

However, extending PE to text is highly non-trivial. PE requires APIs that generate random samples and variations of a given sample, which need to be *redesigned* for text. In particular, unlike generating image variants in the continuous pixel space where diversity can be easily manipulated using existing model hyperparameters (e.g., guidance scale in diffusion model (Ho & Salimans, 2021)), texts operate in a discrete space, making it challenging to effectively *control* the generation diversity. In addition, in contrast to images with fixed dimensionality, text data exhibit varied *lengths* which adds another layer of complexity. To this end, we propose an augmented PE algorithm (AUG-PE) with *new **generation** and **selection** techniques* that allow us to i) elicit a larger set of more diverse and higher-quality texts from LLMs with appropriate sequence length and ii) effectively select the most relevant texts. Our contributions are:

- We propose AUG-PE for high-quality DP synthetic text generation leveraging API access to powerful LLMs. This includes both a practical instantiation of PE on texts and fundamental algorithmic innovations that may benefit future applications of PE.
- We conduct comprehensive evalutions of AUG-PE on Yelp, OpenReview (ICLR 2023), and PubMed (Aug 2023) datasets with various LLMs, including GPT-2-series models, GPT-3.5, and open-source LLMs. We show that under *the same pretrained LM* (GPT-2-series) and privacy budget $\epsilon = 4, 2, 1$, AUG-PE can generate DP synthetic text that achieves comparable or even better performance than finetuning baselines in some cases, in terms of downstream task utility and similarity between synthetic and real samples. Leveraging *more powerful LLMs* such as GPT-3.5 (where DP finetuning is not applicable) and five open-source LLMs (where DP finetuning is hard to implement), the performance of AUG-PE can be significantly improved. Additionally, AUG-PE can be more computationally efficient than DP finetuning by requiring LLM *inference* APIs only.
- We explore the properties of AUG-PE including its text length distribution, its compatibility with stronger LLMs as data generators and downstream models, and its behaviors under data scaling, to provide insights for future development of PE.

2. Background

Differential Privacy (DP). (ϵ, δ)-DP ensures that the output of a randomized mechanism $\mathcal{M}$ is close regardless of whether an individual data record is included in the input or not. Specifically, given any pair of two adjacent datasets $\mathcal{D}, \mathcal{D}'$ (i.e., adding or removing one sample), any possible output set E, it holds that $\Pr[\mathcal{M}(\mathcal{D}) \in E] \leq e^{\epsilon} \Pr\left[\mathcal{M}\left(\mathcal{D}'\right) \in E\right] + \delta$. Moreover, arbitrary post-processing of the output of an (ϵ, δ)-DP mechanism does not incur additional privacy loss, based on the *post-processing property* of DP (Dwork et al., 2014).

DP synthetic text. To guarantee DP for private training data, one method involves using DP-SGD (Abadi et al., 2016) during model training for specific NLP tasks (Yu et al., 2022; Li et al., 2021). Alternatively, one can finetune pretrained generative language models, such as GPT-2, with private data using DP-SGD and then generate synthetic text datasets (Putta et al., 2022; Bommasani et al., 2019)

(Fig. 1). Such DP synthetic texts can be employed in an arbitrary number of non-privately trained downstream tasks without increasing privacy loss. Studies by Yue et al. (2023); Mattern et al. (2022a); Kurakin et al. (2023) indicate that training downstream models on DP synthetic text yields performance akin to directly training them on real data with DP, highlighting the good quality of synthetic data.

However, given that state-of-the-art LLMs (e.g., GPT-4, Claude, GPT-3.5) do not provide model weights, DP finetuning them is infeasible. Even for open-source LLMs (e.g., LLaMA (Touvron et al., 2023a;b)), it is resource-intenstive to perform finetuning (Malladi et al., 2023). Finetuning with DP-SGD is even harder due to the well-known challenges of *per-sample gradient* calculations for clipping to guarantee DP. Even with optimization techniques (Malladi et al., 2023; He et al., 2022), DP finetuning is still memory and computationally intensive due to large batch sizes and long training iterations required to reach a good fidelity-privacy trade-off (Anil et al., 2021). Here, we study an API-based method for DP synthetic text generation to overcome these challenges, which only requires model inference and is applicable no matter whether the LLM is open-sourced or not.

Additionally, there is a line of work on text-to-text privatization techniques, which provide different privacy guarantees than DP, such as word-level metric DP or sample-level local DP. We defer more discussion and comparison to App. C.11.

3. Method

3.1. Preliminaries on Private Evolution (PE)

PE is recently proposed as an alternative to DP finetuning for DP synthetic data generation (Lin et al., 2024) by merely requiring APIs of pretrained models, and thus is easier to implement and deploy and can leverage API-based models. The original PE algorithm (for unconditional generation)[3] is the $L = 1$ case in Alg. 1. PE works by first calling RANDOM_API that generates random samples from the foundation model (Line 2), and then iteratively: (1) using private samples to vote for their nearest synthetic samples (under embedding model Φ) to construct a DP_NN_HISTOGRAM (Line 11), (2) drawing samples according to the histogram (Line 15), and (3) passing those samples through VARIATION_API which generates new samples that are similar to the given one (Line 16), e.g., images with a similar object.

While the PE framework is general across modalities, its core components including Φ (the embedding model), RANDOM_API (API for generating random samples from the pretrained model), and VARIATION_API (API for generating new samples that are similar to the given one) require domain-specific designs, and the original paper (Lin et al., 2024) only explores their implementation for images. Compared to images, text introduces unique challenges. For example, unlike images which have a fixed dimensionality, the length of text can vary. In addition, the original PE algorithm yields unsatisfactory text quality. In the following, we explore our design choices for each component and propose our augmented version on text, Aug-PE (shown in Alg. 1 and Fig. 2) with new algorithmic techniques to increase the diversity and quality of text generation.

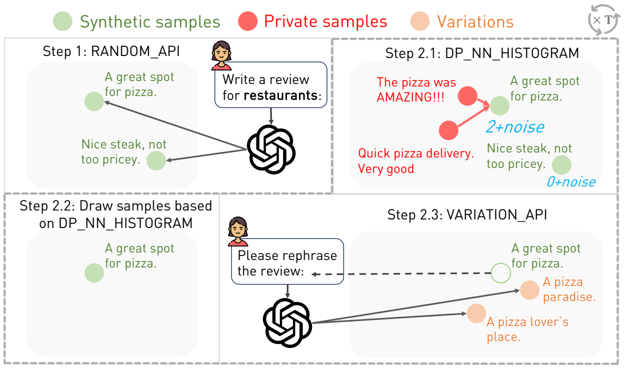

Figure 2: Overview of Aug-PE. We use two private & synthetic samples (reviews for the "restaurant" class) for illustration. **Step 1** (RANDOM_API, Line 2): we use prompts to generate random samples from the LLM. **Step 2**: we iteratively go through steps 2.1-2.3 to refine the synthetic samples towards the private samples. **Step 2.1** (Line 11): each private sample votes for their closet synthetic sample (using self-embedding Line 6 or mean embedding Line 9) in the embedding space induced by embedding model Φ. "A great spot for pizza" gets 2 votes, and the other sample gets 0 votes. We then add Gaussian noise to the votes to ensure DP. This gives us the DP Nearest Neighbor Histogram (DP_NN_HISTOGRAM). **Step 2.2**: we resample the generated texts according to the histogram. We assume that only "A great spot for pizza" remains. **Step 2.3** (VARIATION_API): we use prompts to ask the LLM to generate new similar samples, which are the initial synthetic samples in the next iteration. The prompts are simplified for illustration; see App. B for the complete prompts.

3.2. Aug-PE Design

RANDOM_API. Given the strong instruction-following capability of LLMs, we consider directly using prompts to generate samples (step 1 in Fig. 2). Following Yue et al. (2023), we assume that class labels are non-private. Therefore, we put class label in the prompt (e.g., "restaurant" in Fig. 2). To encourage diverse generation, we propose a *pseudo-class* approach, where we generate a list of subcategories for each class from GPT-3.5 and randomly sample one subcategory as the keyword to put in the prompt for each generation (e.g., *Steakhouse*, *Bistros* for restaurants).

VARIATION_API takes a sample as input and outputs its variations.[4] Unlike image diffusion models used in Lin et al.

[3] The conditional version of PE is running Alg. 1 for the private samples from each class/label separately; see Lin et al. (2024).

[4] While the function processes each sample independently, for notation simplicity, we input an entire dataset to VARIATION_API,

(2024), text models usually do not provide off-the-shelf variation APIs. Again, we leverage the instruction-following capability of LLMs to implement this via prompting. We propose two variation methods: *paraphrasing* and *fill-in-the-blanks*. For *paraphrasing*, we use the prompt "Please rephrase the below sentences: $\{\text{input}\}$". For *fill-in-the-blanks*, we mask $p\%$ tokens of input as blanks, resulting in masked_input, and use "Please fill in the blanks for the below sentences: $\{\text{masked\_input}\}$" as the prompt. Given the in-context learning ability of recent LLMs, we provide *few-shot demonstrations* to improve the generation quality. To add diversity to the generated variations, we create *tone* candidates (e.g., "in a creative way", "in a professional style"), randomly subsample one tone, and add such phrase into the prompt for each generation.

Algorithm 1 Augmented Private Evolution (AUG-PE)

```
Input: private dataset $S_{\text{pri}}$, noise multiplier $\sigma$, text embedding model $\Phi$, number of
       synthetic samples $N_{\text{syn}}$, $K$, $L$
Output: Synthetic text dataset $S_{\text{syn}_T}$
1  $E_{\text{pri}} = \Phi(S_{\text{pri}})$
2  $S_0 \leftarrow$ RANDOM_API $(N_{\text{syn}} *\text{L})$
3  for iteration $t = 0$ to $T - 1$ do
4      // embedding calculation for synthetic samples
5      if $K == 0$ then
6          $E_t = \Phi(S_t)$
7      else if $K > 0$ then
8          $S_t^k \leftarrow$ VARIATION_API $(S_t)$ for $k = 1, 2 \ldots, K$
9          $E_t = \frac{1}{K}\sum_{k=1}^{K}\Phi(S_t^k)$
10     // DP histogram calculation
11     $\text{Histogram}_t \leftarrow \texttt{DP\_NN\_HISTOGRAM}(E_t, E_{\text{pri}}, \sigma)$
12     $P_t \leftarrow \text{Histogram}_t$ / sum $(\text{Histogram}_t)$
13     // synthetic sample selection and generation
14     if $L == 1$ then
15         $S'_t \leftarrow$ draw $N_{\text{syn}}$ samples with replacement from $S_t$ with probability $P_t$
16         $S_{t+1} \leftarrow$ VARIATION_API $(S'_t)$
17         save dataset $S_{\text{syn}_{t+1}} \leftarrow S_{t+1}$
18     else if $L > 1$ then
19         $S'_t \leftarrow$ rank samples by probabilities $P_t$ and draw top $N_{\text{syn}}$ samples
20         save dataset $S_{\text{syn}_{t+1}} \leftarrow S'_t$
21         $S_{t+1}^j \leftarrow \text{VARIATION\_API}(S'_t)$ for $\mathbf{j = 1, 2 \ldots, L - 1}$
22         $S_{t+1} \leftarrow [S_{t+1}^1, \ldots, S_{t+1}^{L-1}, \mathbf{S'_t}]$
23
24 return $S_{\text{syn}_T}$
25 Procedure DP_NN_HISTOGRAM$(E_{\text{syn}}, E_{\text{pri}}, \sigma)$
       Input: synthetic embedding set $E_{\text{syn}} = \{e_j\}_{j=1}^n$, private embedding set $E_{\text{pri}}$,
              noise level $\sigma$, distance function $d(\cdot, \cdot)$
26     $\text{Histogram} \leftarrow [0, \ldots, 0]$
27     for $e_{\text{pri}} \in E_{\text{pri}}$ do
28         $i = \arg\min_{j \in [n]} d(e_{\text{pri}}, e_j)$;
29         $\text{Histogram}[i] \leftarrow \text{Histogram}[i] + 1$
30     $\text{Histogram} \leftarrow \text{Histogram} + \mathcal{N}(0, \sigma^2 I_n)$
31     return Histogram
```

Adaptive text lengths in VARIATION_API. The distribution of text length in real-world datasets is usually fat-tailed: most samples are short while a few are long (Fig. 4). In DP-finetuning-based approaches, to faithfully capture long texts, we need to set a large max token length (denoted by max_token). However, this would significantly increase the computation cost. Prior work (Yue et al., 2023) circumvents this problem by setting a *small* max_token at the cost of the capability to generate long texts. AUG-PE faces the same challenge. Since APIs usually charge by token usage, a high max_token raises costs (as generated text can exceed needs), while a low max_token sacrifices fidelity.

To address the challenge, we leverage PE to learn text lengths automatically by adjusting *per-sample* max_token adaptively. Specifically, in VARIATION_API, we add "with $\{\text{targeted\_word}\}$ words" in the prompt to specify the desired word count in the generation. targeted_word is modified by setting $\text{targeted\_word} = \max\{\text{original\_word} + \mathcal{N}(0, \sigma_{word}^2), \text{min\_word}\}$ where original_word is the word count of input, σ_{word}^2 is Gaussian noise variance and min_word is a minimal targeted word ensuring useful generations. We set $\text{max\_token} = \lfloor \text{targeted\_word} * \text{w2t\_ratio} \rfloor$ for LLM API calls where w2t_ratio is the approximate number of tokens per word (OpenAI, 2023c).

Embeddings calculation and DP nearest neighbor histogram. We use off-the-shelf text embedding models Φ to calculate the embedding of private/synthetic samples. Notably, the embedding of synthetic samples can be defined either by their self-embedding (when $K = 0$) or the averaged embedding from K variations (when $K > 0$). After calculating embeddings, each private sample votes for its nearest synthetic sample in the embedding distance, which results in the Histogram_t for synthetic samples. As the voting utilizes private samples, we add Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each bin of Histogram_t to ensure DP.

Sample selection and generation. AUG-PE introduces significant enhancements over the original PE for generating more diverse samples and selecting/retaining high-quality samples. Specifically, to *enhance sample diversity*, we propose the following methods: (1) The random sampling based on the histogram probability P_t (Line 15) in original PE results in repeated samples, causing performance degradation for S'_t. To mitigate this, AUG-PE ranks synthetic samples according to their probability and selects only the top N_{syn} samples, enhancing the diversity without sample redundancy (Line 19). (2) Instead of a single variation, AUG-PE generates $L - 1$ variations for each selected sample in S'_t, creating a larger and more diverse synthetic dataset S_{t+1} for subsequent iterations (Line 21). (3) We modify the size of the initial dataset to be L times larger than N_{syn}, matching the expanded size of S_{t+1} (Line 2). To *select/retain high-quality samples*, we propose the following methods: (1) The selected samples S'_t are also included in the next iteration's dataset S_{t+1}, increasing the likelihood of retaining high-quality synthetic candidates (Line 22). (2) For LLMs, we find that when the variation API produces samples with large variations, the averaged embedding from the variations is not representative of the actual sample. Therefore, we use $K = 0$ so the nearest neighbor voting is

which outputs corresponding variations for each sample within it.

performed on the self-embedding of synthetic samples and we directly use those selected, good samples as algorithm's output $S_{\text{syn}_{t+1}} \leftarrow S'_t$ (Line 20).

In practice, we use $\{K = \#\text{variations}, L = 1\}$ as original PE, and $\{K = 0, L = \#\text{variations} + 1\}$ as AUG-PE, so that the number of API calls for generating variations (i.e., #variations) are kept the same for fair comparisons.

These enhancements position AUG-PE as a more effective method to generate diverse and high-quality synthetic *text*.

Privacy analysis of AUG-PE follows original PE and we provide detailed privacy analysis in App. A. Specifically, since each private sample only contributes 1 vote for one bin in the histogram (i.e., nearest synthetic sample), the sensitivity is 1. The histograms are privatized by adding Gaussian noise. The adaptive DP composition theorem (Dong et al., 2019) is applied to track the privacy loss across T iterations.

4. Experiments

Datasets. We evaluate AUG-PE on three datasets: Yelp Review (Inc, 2023), OpenReview, and PubMed abstracts. We use Yelp, a public benchmark providing reviews on businesses, following the choice in prior work for DP synthetic text (Yu et al., 2022). To mitigate the concerns that existing benchmarks are potentially used at LLM's pretraining stage, we crawl the latest reviews for ICLR 2023 submissions from OpenReview website[5] to construct a new dataset, where the reviews are made public after recent LLMs are trained. We also use PubMed with abstracts of medical papers[6] crawled by Yu et al. (2023) from 2023/08/01 to 2023/08/07 after recent LLMs are trained. Notably, texts from Yelp are mainly in styles of daily conversation, while the other two datasets require domain-specific knowledge about machine learning or biomedical literature when generating DP synthetic replicas. For conditional generation, we use below attributes as labels: the review ratings and business category for Yelp, and the review recommendation and area for OpenReview. For PubMed, we use unconditional generation.

Models. For data generators, we use GPT-2 (Radford et al., 2019), GPT-2-Medium, GPT-2-Large, GPT-3.5 (OpenAI, 2022), and non-GPT based LLMs including four 7b-sized models – OPT (Zhang et al., 2022), Vicuna (Zheng et al., 2023), Falcon (Almazrouei et al., 2023), LLaMA-2 – as well as one Mixture-of-Expert model Mixtral-8x7B (MistralAI, 2022). For embedding models, we use sentence-transformer (Reimers & Gurevych, 2019). *We study more types of embedding models as ablation study in § 4.2.*

Metrics. We evaluate synthetic texts regarding (i) accuracy on downstream tasks, and (ii) similarity between real and synthetic data.

Downstream tasks: we finetune downstream models on the synthetic text and evaluate their accuracy on the real test dataset. We pick two representative use cases: using DP synthetic text to train DP text classifiers (Yue et al., 2023) and to train efficient DP lanaguge models (Yu et al., 2023). Specifically, we finetune RoBERTa-base (Liu et al., 2019) as text classifiers to classify review ratings and business categories for Yelp, and to classify review recommendations and areas for OpenReview. For PubMed, we finetune $\text{BERT}_{\text{Mini}}$/$\text{BERT}_{\text{Small}}$ (Turc et al., 2019)[7] on synthetic text and evaluate their next-word prediction accuracy. *We study more types downstream models as ablation study in § 4.2.*

Similarity between real and synthetic data: we quantitively compare (a) *embedding distribution distance* (i.e., Fréchet Inception Distance (FID) (Heusel et al., 2017), Precision, Recall, F1 score (Kynkäänniemi et al., 2019), MAUVE score (Pillutla et al., 2021), KL and TV divergences (Chung et al., 1989)) and qualitatively compare (b) *text length distribution difference* (Yue et al., 2023).

Baselines. We consider two SOTA baselines involving DP finetuning: (1) DP-FT-DOWNSTREAM (Yu et al., 2022; Li et al., 2022): finetuning downstream model on real data with DP-SGD. Note that this baseline is not a competitor to our method, since *our goal is to generate DP synthetic data and not merely train a downstream model.* (2) DP-FT-GENERATOR (Yue et al., 2023): finetuning generator (e.g., GPT-2) with DP-SGD (note that we cannot finetune closed-source GPT-3.5) and using synthetic texts to finetune downstream model with non-private SGD.

We defer more details about the setups, hyperparameters and metrics to App. B.

4.1. Understanding the Performance of AUG-PE

Here, we analyze the performance of AUG-PE by answering five research questions about its utility, efficiency, and robustness against empirical privacy attacks under DP compared to DP-finetuning-based baselines.

RQ1: Can DP synthetic texts generated from AUG-PE *outperform those from* DP-FT-GENERATOR*?* **DP synthetic texts from AUG-PE can have comparable privacy-utility trade-off to those from DP-FT-GENERATOR using the same generator, while outperforming it using the stronger generator GPT-3.5.** The downstream model accuracy of different methods along 4 generators on different benchmark datasets is shown in Tb. 1. **(1)** When using the same LM (GPT-2-series) as the generator for fair compar-

[5] https://openreview.net/group?id=ICLR.cc/2023/Conference

[6] https://www.ncbi.nlm.nih.gov/

[7] Following (Yu et al., 2023), we apply a causal language modeling mask that restricts each token to only attend to its preceding tokens.

Table 1: Evaluation on downstream model accuracy of three methods along 4 data generators. The highest accuracy across all methods (obtained by AUG-PE) is **bolded** (underlined). (i) Compared to DP-FT-GENERATOR, in some cases, downstream accuracy of AUG-PE is higher (↑) under the same size of GPT-2-series data generator. Leveraging the inherent knowledge within stronger LLM, GPT-3.5, AUG-PE can achieve higher accuracy, especially on challenging datasets OpenReview and PubMed, outperforming DP-FT-GENERATOR by a notable margin. (ii) Compared to traditional method DP-FT-DOWNSTREAM, AUG-PE can also obtain higher accuracy under DP.

Dataset	**Method**	**Data Type (Size)**	**Data Generator**	$\epsilon = \infty$		$\epsilon = 4$		$\epsilon = 2$		$\epsilon = 1$	
				Rating	Category	Rating	Category	Rating	Category	Rating	Category
Yelp	DP-FT-DOWNSTREAM	Original (1939290 / full data)	-	**76.0**	**81.6**	67.5	72.8	67.2	72.0	66.8	71.8
		Original (5000)		70.5	75.1	44.8	61.8	44.8	61.8	44.8	61.8
	DP-FT-GENERATOR	Synthetic (5000)	GPT-2	70.3	75.9	68.2	74.1	67.2	73.1	66.4	73.9
			GPT-2-Medium	70.0	75.0	**69.0**	74.6	67.8	74.3	67.4	74.1
			GPT-2-Large	70.4	75.4	68.7	74.2	**69.8**	**75.1**	**68.7**	74.6
	AUG-PE	Synthetic (5000)	GPT-2	67.5	74.8	66.4	**74.9** ↑	67.1	74.7 ↑	66.9 ↑	74.4 ↑
			GPT-2-Medium	67.5	74.9	66.8	74.6	67.7	74.7 ↑	67.3	74.6 ↑
			GPT-2-Large	67.5	74.5	67.3	74.4 ↑	65.8	74.1	66.5	**75.0** ↑
			GPT-3.5	68.4	74.1	68.1	74.0	67.8	74.3	67.9	74.0
				Area	Rating	Area	Rating	Area	Rating	Area	Rating
OpenReview	DP-FT-DOWNSTREAM	Original (8396 / full data)	-	**65.1**	**50.8**	30.5	32.0	30.5	32.0	30.5	32.0
		Original (2000)		55.3	47.8	30.5	32.0	30.4	25.5	6.3	19.8
	DP-FT-GENERATOR	Synthetic (2000)	GPT-2	47.5	32.0	32.1	32.0	31.9	32.0	32.1	32.0
			GPT-2-Medium	49.7	36.5	40.3	32.0	33.5	31.9	35.5	31.9
			GPT-2-Large	48.3	42.9	38.9	33.7	40.4	33.6	38.6	32.1
	AUG-PE	Synthetic (2000)	GPT-2	42.4	32.1 ↑	39.9 ↑	32.1 ↑	38.8 ↑	32.1 ↑	37.6 ↑	32.0
			GPT-2-Medium	41.0	32.3	36.9	32.0	36.0 ↑	32.0 ↑	36.6 ↑	32.1 ↑
			GPT-2-Large	42.1	32.1	38.8	32.0	38.4	32.0	38.1	32.0
			GPT-3.5	45.4	43.5	**43.5**	**44.6**	**42.8**	**44.5**	**41.9**	**43.1**
				$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
PubMed	DP-FT-DOWNSTREAM	Original (75316 / full data)	-	**43.5**	**47.6**	**30.7**	**34.1**	28.9	**32.5**	26.7	30.4
		Original (2000)		33.5	34.6	2.2	1.1	1.8	0.8	1.4	0.6
	DP-FT-GENERATOR	Synthetic (2000)	GPT-2	30.2	32.4	27.8	29.7	27.6	29.3	27.2	29.2
			GPT-2-Medium	31.0	33.1	28.4	30.2	28.1	30.0	27.8	29.8
			GPT-2-Large	31.0	33.1	29.2	31.2	29.2	31.1	28.9	31.1
	AUG-PE	Synthetic (2000)	GPT-2	24.5	26.7	24.7	27.0	24.7	26.9	24.3	26.5
			GPT-2-Medium	25.5	27.7	25.4	27.6	25.1	27.4	24.9	27.0
			GPT-2-Large	25.7	28.0	25.8	27.9	25.5	27.7	25.1	27.2
			GPT-3.5	30.4	32.7	30.3	32.5	**30.2**	**32.5**	**30.1**	**32.4**

isons, DP synthetic texts from AUG-PE demonstrate competitive or even better (↑) utility than DP-FT-GENERATOR on Yelp and OpenReview. However, AUG-PE underperforms DP-FT-GENERATOR on PubMed. This is expected because AUG-PE relies on the knowledge within LLMs to generate high-quality texts without domain-specific finetuning, while GPT-2-series models might have limited exposure to biomedical literature (Radford et al., 2019). **(2)** AUG-PE only requires API access, making it possible to use closed-source LLM such as GPT-3.5 for generating DP synthetic text. The results of GPT-3.5 outperform not only AUG-PE GPT-2-series, but also DP-FT-GENERATOR GPT-2-series by a significant margin, especially on challenging datasets such as OpenReview and PubMed. It shows that AUG-PE can effectively leverage the inherent knowledge (e.g., medical knowledge, sentiment of reviews, research areas about machine learning) in stronger LLMs to generate higher-quality DP synthetic texts. **(3)** In addition to downstream utility, we measure the *embedding distribution distance* between real and synthetic samples. The results in App. C.9 show that AUG-PE can obtain similar and even lower distances (reflected by FID, TV divergence, Recall, F1, and MAUVE scores, etc.) compared to DP-FT-GENERATOR. **(4)** Some methods consistently show a 32.0 accuracy for Rating and 30.5 for Area classification, due to the failure of the downstream RoBERTa-base model under DP, always outputting majority class (see App. B for label distributions).

RQ2: Can DP synthetic texts from AUG-PE *be a better choice than* DP-FT-DOWNSTREAM *on real data with DP?* **AUG-PE obtains comparable and higher accuracy than DP-FT-DOWNSTREAM under DP. (1)** Tb. 1 shows that under $\epsilon = 2, 1$ on PubMed, AUG-PE GPT-3.5 with a smaller synthetic dataset size (2k) is sufficient to produce better downstream models compared to models directly trained with DP on the original data of the full (75k) or same size (2k). Similar conclusions hold for other two datasets, and the advantages of AUG-PE on OpenReview are evident across all generators. **(2)** DP-FT-DOWNSTREAM performs fairly poor when the data size is small (e.g., 2k on PubMed and OpenReview), indicating that LMs finetuned with DP-SGD is unable to learn meaningful information under DP noises when samples are limited (Yu et al., 2021; Li et al., 2022; Bu et al., 2022). In contrast, postprocessing property of DP allows us to train downstream tasks on DP synthetic text (with any size) via normal training techniques, without incurring additional privacy loss, potentially leading to a better downstream model than DP-FT-DOWNSTREAM.

RQ 3: How does AUG-PE *perform across different privacy budget ϵ?* **(1)** Tb. 1 shows that AUG-PE in general achieves better performance as ϵ increases from $1, 2, 4$ to ∞, suggesting that AUG-PE scales well with the privacy

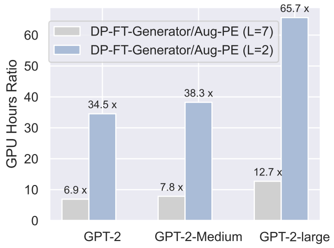

Figure 3: Efficiency comparison between DP-FT-GENERATOR and AUG-PE on Yelp for generating 100k synthetic samples ($\epsilon = 1$)

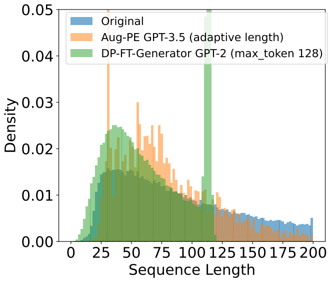

Figure 4: GPT-3.5 with adaptive text length achieves a comparable text length distribution to the original data on Yelp.

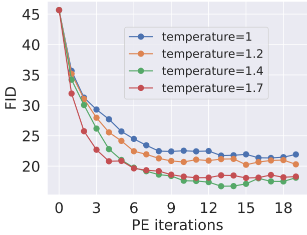

Figure 5: Larger temperature for GPT-3.5 leads more diverse generation on Yelp with a lower FID score.

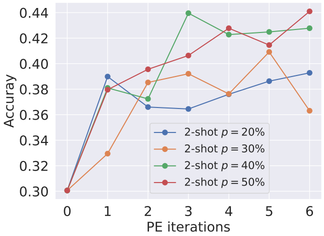

Figure 6: Fill-in-the-blanks with a larger mask probability $p\%$ for GPT-3.5 leads to more diverse generation and higher utility on OpenReview.

budget ϵ. **(2)** On OpenReview, from $\epsilon = \infty \rightarrow 1$, the rating classification accuracy obtained from DP-FT-GENERATOR GPT-2-Large generated text drops from $48.3 \rightarrow 38.6$, and DP-FT-DOWNSTREAM on full training data drops from $65.1 \rightarrow 30.5$, while the accuracy of AUG-PE GPT-3.5 exhibits marginal drop $45.4 \rightarrow 43.1$. It suggests that in some cases, **the performance of AUG-PE (paired with powerful generator) can be more robust under DP noise than FT baselines**. The reason could be that LMs are vulnerable to the perturbations introduced in model parameters through DP-SGD, whereas AUG-PE strategically adds noise to the histogram votes, effectively preserving the utility.

RQ 4: Compared to DP-FT-GENERATOR*, how efficient the API-access-based* AUG-PE *is in terms of GPU hours?* **With inference API access, AUG-PE is more efficient than DP-FT-GENERATOR that requires DP-SGD finetuning**. **(1)** As shown in Fig. 3, to generate 100k synthetic samples on Yelp under $\epsilon = 1$, given the same generator GPT-2-Large, AUG-PE $L = 7$ provides 12.7x speedup and $L = 2$ further provides 65.7x speedup. **(2)** The running time of AUG-PE is mainly scaled with # API calls, which is associated with the number of variations $L - 1$ in Line 21. **(3)** The bottleneck of DP-FT-GENERATOR is DP-SGD finetuning: it takes 1764 GPU hours on 32G NVIDIA V100 to finetune GPT-2-Large on Yelp and 7 hours to generate 100k samples, while AUG-PE $L = 2$ ($L = 7$) only requires 27 hours (139 hours). It highlights the computational expense of DP-SGD training, particularly for training LLMs, and underscores the efficiency of the API-based DP algorithm AUG-PE. A detailed breakdown of the GPU hours for each setting is in Appendix Tb. 23. **(4)** We use half precision (FP16) for LLM inference in AUG-PE. With the emerging efficient inference techniques (e.g., Liu et al. (2023)), AUG-PE runtime can be further optimized.

RQ 5: How robust AUG-PE *is under empirical privacy attacks compared to DP-finetuning-based baselines?* We perform state-of-the-art text membership inference attacks (MIAs) against the finetuned downstream models on PubMed dataset. We consider three types of MIAs and report the AUC score: (1) *PPL* thresholds perplexity to predict membership (Carlini et al., 2021); (2) *REFER* computes the ratio of the log perplexity of the tested model against a reference model (Carlini et al., 2021); (3) *LIRA* uses the ratio of likelihood (Carlini et al., 2022) and we use the pre-trained model as a reference following (Mattern et al., 2023). The results in Tb. 3 show that **AUG-PE exhibits lower AUC scores under MIAs compared to DP-FT-GENERATOR and DP-FT-DOWNSTREAM**. This indicates a higher robustness to empirical privacy attacks, potentially due to the synthetic nature of the data used for downstream model finetuning, which inherently reduces the risk of overfitting to real private data. We defer the details to App. C.1.

4.2. Understanding the Properties of AUG-PE

Here we study properties of AUG-PE including text lengths, its compatibility with stronger data generators and downstream models, and its behaviors under data scaling.

RQ 6: Can AUG-PE *produce sentence length distributions similar to real data?* **AUG-PE produces favorable text length distributions.** From Fig. 4, we see that the text length distribution of synthetic samples produced from GPT-3.5 through AUG-PE is close to the distribution of the original Yelp data, highlighting the effectiveness of our adaptive sequence length mechanism (§ 3.2). Note that the finetuning baseline requires a fixed max_token (e.g., 128 for GPT-2), which leads to a hard threshold for maximal text length, which is not the case in our method with our adaptive length technique. Nevertheless, there is a peak near 30 tokens for AUG-PE, which is due to the min _word set in the prompt to prevent empty generation. We defer the convergence of text length distributions over PE iterations to App. C.2.

RQ 7: Can AUG-PE *benefit from more powerful LLMs?* **AUG-PE is effective across a wide range of API-accessible LLMs.** We have observed from Tb. 1 that GPT-3.5 can lead to higher downstream accuracy than GPT-2-series, especially on PubMed and OpenReview. Here we evaluate more API-accessible, non-GPT based LLMs. **(1)** As shown in Tb. 2, under $\epsilon = \infty, 1$, those modern LLMs can obtain comparable and even higher accuracy than GPT-3.5 on Yelp, suggesting that AUG-PE can effectively elicit and select high-quality synthetic text from various types of LLMs. Note that *DP finetuning often needs to be imple-*

Table 2: Using powerful LLMs as data generators leads to improved downstream accuracy on three datasets.

	Yelp				OpenReview				PubMed			
	$\epsilon = \infty$		$\epsilon = 1$		$\epsilon = \infty$		$\epsilon = 1$		$\epsilon = \infty$		$\epsilon = 1$	
LLM	Rating	Category	Rating	Category	Area	Rating	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
GPT-2	67.5	74.8	66.9	74.4	42.4	32.1	37.6	32.0	24.5	26.7	24.3	26.5
GPT-2-Medium	67.5	74.9	67.4	74.6	41.0	32.3	36.6	32.1	25.5	27.7	24.9	27.0
GPT-2-Large	67.5	74.5	66.6	75.0	42.1	32.1	38.1	32.0	25.7	27.9	25.1	27.2
Opt-6.7b	68.7	**75.3**	67.7	**75.3**	43.6	32.2	30.5	32.1	26.5	28.6	25.8	27.9
Vicuna-7b-v1.5	**68.8**	74.1	67.2	74.9	42.9	35.7	35.2	35.4	24.6	26.9	23.1	24.9
Falcon-7b-instruct	67.4	74.9	67.3	74.2	38.6	32.6	39.0	33.3	22.3	24.4	22.4	24.5
Llama-2-7b-chat-hf	68.6	74.9	**68.0**	75.1	45.5	38.5	36.4	37.0	25.8	28.4	24.8	27.5
Mixtral-8x7B-v0.1	68.2	74.6	67.6	74.6	**45.9**	41.8	**43.6**	42.3	24.9	27.6	24.5	27.1
GPT-3.5	68.4	74.1	67.9	74.0	45.4	**43.5**	41.9	**43.1**	**30.4**	**32.7**	**30.1**	**32.4**

Table 3: Aug-PE generally yields lower AUC scores against membership inference attacks on PubMed than DP-FT-Generator and DP-FT-Downstream, indicating a higher robustness to empirical privacy attacks.

Method	**Generator**	**AUC ($\epsilon = \infty$)**			**AUC ($\epsilon = 4$)**			**AUC ($\epsilon = 2$)**			**Avg**
		PPL	REFER	LIRA	PPL	REFER	LIRA	PPL	REFER	LIRA	
DP-FT-Downstream	/	77.60	74.93	65.05	**49.32**	54.58	56.82	**48.96**	50.41	51.56	58.80
DP-FT-Generator	GPT-2	55.50	51.35	51.97	53.90	51.12	51.84	53.31	50.81	51.61	52.38
DP-FT-Generator	GPT2-M	54.91	51.25	51.88	54.72	51.13	51.76	54.58	51.28	51.85	52.60
DP-FT-Generator	GPT2-L	54.81	51.22	51.86	54.56	50.81	51.64	55.05	51.01	51.69	52.52
Aug-PE	GPT-2	50.08	50.92	51.66	50.10	50.97	51.70	49.94	50.85	51.64	50.87
Aug-PE	GPT2-M	49.85	50.73	51.57	50.10	50.95	51.69	49.73	50.65	51.51	50.75
Aug-PE	GPT2-L	**49.43**	50.40	51.40	49.61	**50.56**	51.48	49.66	50.60	51.49	**50.51**
Aug-PE	GPT-3.5	52.23	**49.67**	**50.85**	52.68	49.84	**50.93**	52.77	**49.77**	**50.85**	51.07

Table 4: The next word prediction accuracy increases when using larger downstream models for PubMed synthetic texts.

ϵ	Method	Generator	bert-tiny 4.4M	bert-mini 11.2M	bert-small 28.8M	Llama2-7b-chat-hf 7B
∞	DP-FT-Generator	GPT-2-Large	**24.6**	**31.0**	**33.1**	53.1
	Aug-PE	GPT-3.5	23.0	30.3	32.7	**56.5**
1	DP-FT-Generator	GPT-2-Large	**23.1**	28.9	31.1	52.0
	Aug-PE	GPT-3.5	22.9	**30.1**	**32.4**	**56.4**

mented case-by-case for LLMs and currently lacks opensource implementations for these LLMs, whereas Aug-PE can easily leverage them. **(2)** The results on OpenReview and PubMed in Tb. 2 show that GPT-3.5 leads to higher utility than opensource LLMs (e.g. LLaMA-2), demonstrating the stronger generation power of GPT-3.5 in academic/medical domains. Interestingly, Mixtral-8x7B can also generate high-quality synthetic texts for OpenReview, but not for PubMed.

RQ 8: Can more powerful downstream models benefit from synthetic text generated via Aug-PE*?* **The high-quality synthetic text from Aug-PE is better utilized by larger downstream models.** **(1)** From each row in Tb. 4, we see that next-word prediction accuracy monotonically increases with the use of larger downstream models trained on PubMed synthetic text. **(2)** Under both $\epsilon = 1, \infty$, the smallest model $\text{BERT}_{\text{Tiny}}$ favors the synthetic texts from DP-FT-Generator GPT-2-Large, while larger models such as LLaMA-2 favor synthetic text from Aug-PE GPT-3.5. This observation underscores the importance of choosing downstream models of a suitable size; employing overly small models could under-estimate the quality of synthetic texts produced by Aug-PE with GPT-3.5. We hypothesize that this is because i) GPT-3.5 generated texts might already be of higher quality in terms of vocabulary, syntax, semantic coherence, etc., compared to generated texts from finetuned GPT-2-Large; and ii) larger downstream LMs like LLaMA-2 can better understand and utilize the nuances in synthetic texts for improved performance than $\text{BERT}_{\text{Tiny}}$.

RQ 9: Can we further improve downstream task accuracy with more synthetic samples generated from Aug-PE*?* To study the scaling law of Aug-PE, we use GPT-2-series models to generate {5k,10k,100k} samples for Yelp, and {2k,3k,5k} samples for other two datasets. As shown in App. C.10, under $\epsilon = 1, 2, 4, \infty$, Aug-PE in general achieves better performance across all datasets as the data size increases, suggesting that **Aug-PE scales well with the number of synthetic samples.**

4.3. Validating the Design of Aug-PE

As Aug-PE introduces novel sample selection and generation techniques, here we study algorithm components related to the two steps, respectively (under $\epsilon = \infty$), and compare its performance against the original PE.

RQ 10: Can Aug-PE *surpass original* PE*?* Tb. 5 shows that **Aug-PE achieves notable improvement over PE** for GPT-2, e.g., +22.6% on Yelp rating classification. We observe similar conclusions for GPT-3.5 in Tb. 24 in App.

Table 5: Aug-PE outperforms PE with GPT-2 on all datasets.

Method	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
PE ← Aug-PE ($k = 6, L = 1$)	44.9	71.8	35.3	32.0	20.1	22.3
Aug-PE ($k = 0, L = 7$)	**67.5**	**74.8**	**42.4**	**32.1**	**24.5**	**26.7**

RQ 11: How does the private data guided sample selection affect Aug-PE *performance?* Here we aim to verify the **components related to sample selection: i) usage of private data; ii) rank-based selection; iii) embedding model** used during nearest neighbor voting.

i) Usage of private data. Tb. 6 shows that the initial samples (generated from Random API) or their variants (generated from Random API + Variation API) exhibit limited utility without using private data. However, the quality of the synthetic text improves notably after just one iteration of Aug-PE ($t = 1$) when guided by private data, and this improvement continues to amplify with T iterations. We report the results under DP in App. C.8.

ii) Rank-based sampling. The results in App. C.7 indicate

Table 6: Private-data guided sample selection in Aug-PE improves the utility of GPT-3.5 generated texts.

Setting	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
Random API	62.3	73.7	34.4	42.0	29.7	31.9
Random API + Variation API	62.3	73.7	36.4	42.0	29.6	31.9
Aug-PE ($t = 1$)	64.9	73.8	39.3	42.5	30.0	32.2
Aug-PE ($t = T$)	**68.4**	**74.1**	**45.4**	**43.5**	**30.4**	**32.7**

that our proposed rank-based sampling (Line 19) consistently outperforms probability-based random sampling in the original PE (Line 15), due to the elimination of sample redundancy inherent in random sampling, as rank-based sampling exclusively selects the top N_{syn} samples.

iii) Embedding models. Tb. 7 shows that larger embedding models such as "sentence-t5-xl" can more accurately capture the nuances of texts in the embedding space, leading to higher utility for GPT-2 generated texts.

Table 7: More powerful embedding model leads to higher utility for GPT-2 generated texts via Aug-PE.

Embeddding model (Reimers & Gurevych, 2019)	Yelp		PubMed	
	Rating	Category	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
sentence-t5-xl	**67.6**	75.1	**25.1**	**27.4**
sentence-t5-base	67.2	75.2	24.5	26.7
stsb-roberta-base-v2	67.5	74.8	23.9	26.1
all-MiniLM-L6-v2	62.6	**75.3**	24.7	26.7
paraphrase-MiniLM-L6-v2	64.7	75.1	24.3	26.5
all-mpnet-base-v2	64.1	74.6	24.0	26.0

RQ 11: How to improve the generation quality through Variation API in Aug-PE*?* We analyze key **componenents related to generation: i) variation API prompt designs; ii) LLMs generation configuration (e.g., temperature); iii) number of variations $L - 1$.**

i) Variation API prompt designs. We evaluate the impact of four types of Variation API prompts on Yelp: paraphrasing and fill-in-the-blanks prompts under zero-shot and few-shot settings. **(1)** Qualitatively, we observed that GPT-2 struggles to adhere to the fill-in-the-blanks instruction, often leaving blanks ("__") in the generated texts. In contrast, GPT-3.5 can effectively fill in the blanks, potentially because GPT-3.5 has been instruction-tuned (Wei et al., 2021) and thus follows the instructions better. **(2)** The quantitative results in Appendix Tb. 25 reveal that paraphrasing can be an effective strategy for GPT-2, while fill-in-the-blanks yields better results for GPT-3.5. **(3)** Fill-in-the-blanks offers more control over the diversity of generated content. By increasing the mask probability $p\%$, we can create more room for imaginative responses from GPT-3.5, leading to more diverse generations. As indicated in Fig. 6, a higher mask probability corresponds to increased accuracy in downstream area classification tasks when using GPT-3.5.

ii) Temperature is a key parameter in controlling the diversity of LLM generation. A higher temperature leads LLMs to generate less frequent tokens, thereby increasing diversity. However, an excessively high temperature may result in overly random outputs and potentially hurt generation.

Table 8: For GPT-2 generated texts, high temperatures are preferred for Yelp while moderate temperatures are favored for OpenReview and PubMed to balance generation diversity and quality.

Temperature	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
0.8	66.9	74.2	42.0	**32.2**	24.5	**26.8**
1.0	66.8	74.8	41.5	32.1	**24.5**	26.7
1.2	67.0	74.9	**42.4**	32.1	24.4	26.5
1.4	**67.5**	74.8	40.8	32.0	23.6	25.6
1.7	67.1	**75.2**	40.6	32.1	21.9	24.0

The impact of different temperatures for Aug-PE on GPT-2 is shown in Tb. 8. **(1)** On Yelp, a higher temperature (1.4 to 1.7) proves beneficial for GPT-2, as business reviews often encompass daily conversations with a variety of sentence formats and tones. Additional findings in Fig. 5 indicate that large temperatures can also lead to low (better) FID scores for GPT-3.5. **(2)** Conversely, on OpenReview and PubMed, a moderate temperature setting (around 1.0) is more suitable for GPT-2, as academic and medical literature demand more precise and accurate text generation.

Table 9: Increasing the number of variations $L - 1$ in Aug-PE yields higher utility for GPT-2 generated texts.

$L - 1$	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
1	65.8	74.4	39.2	**32.1**	23.9	26.1
3	66.7	**75.1**	41.1	32.0	24.6	26.8
6	67.5	74.8	42.4	**32.1**	24.5	26.7
9	**67.7**	74.9	**42.7**	32.0	**24.9**	**26.8**

iii) Increasing the number of variations $L - 1$ generally enhances performance of Aug-PE as shown in Tb. 9, due to the expansion of the candidate synthetic sample pool, which increases the likelihood of getting high-quality texts. However, generating more variations requires additional API calls, leading to increased computational costs as discussed in Fig. 3. To balance the trade-off between utility and efficiency, we use $L = 7$ for GPT-2-series experiments.

Aug-PE convergence. We provide generation results showing the convergence of Aug-PE under *one private sample* in App. C.12, which demonstrate our sample selection and generation process in a more direct manner.

5. Conclusion

In this work, we propose Aug-PE for DP synthetic text generation without model training. We conduct comprehensive experiments on three datasets and show that Aug-PE can generate high-quality DP synthetic text with comparable privacy-utility tradeoff to DP finetuning baselines under the same data generator. Leveraging more powerful open-source LLMs or API-based LLMs as data generators, Aug-PE can generate DP synthetic text with improved utility.

Acknowledgements

The authors thank Janardhan Kulkarni and anonymous reviewers for their valuable feedback and suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Anil, R., Ghazi, B., Gupta, V., Kumar, R., and Manurangsi, P. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*, 2021.

Balle, B. and Wang, Y.-X. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Bommasani, R., Wu, S., and Schofield, X. Towards private synthetic text generation. In *NeurIPS 2019 Machine Learning with Guarantees Workshop*, 2019.

Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. Differentially private bias-term only fine-tuning of foundation models. *arXiv preprint arXiv:2210.00036*, 2022.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. *USENIX Security Symposium*, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Carvalho, R. S., Vasiloudis, T., Feyisetan, O., and Wang, K. Tem: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pp. 883–890. SIAM, 2023.

Chew, H. S. J. The use of artificial intelligence–based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. *JMIR Medical Informatics*, 10(4):e32578, 2022.

Chung, J., Kannappan, P., Ng, C. T., and Sahoo, P. Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, 138(1):280–292, 1989.

CNN. Microsoft is bringing chatgpt technology to word, excel and outlook, 2023. URL https://www.cnn.com/2023/03/16/tech/openai-gpt-microsoft-365/index.html.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

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

Feldman, V., McMillan, A., and Talwar, K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 954–964. IEEE, 2022.

Feyisetan, O., Balle, B., Drake, T., and Diethe, T. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pp. 178–186, 2020.

He, J., Li, X., Yu, D., Zhang, H., Kulkarni, J., Lee, Y. T., Backurs, A., Yu, N., and Bian, J. Exploring the limits of differentially private deep learning with group-wise clipping. *arXiv preprint arXiv:2212.01539*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Inc, Y. Yelp dataset, 2023. URL https://www.yelp.com/dataset.

Kurakin, A., Ponomareva, N., Syed, U., MacDermed, L., and Terzis, A. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Lambert, N., Castricato, L., von Werra, L., and Havrilla, A. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. https://huggingface.co/blog/rlhf.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2021.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. *International Conference on Learning Representations*, 2022.

Lin, Z., Gopi, S., Kulkarni, J., Nori, H., and Yekhanin, S. Differentially private synthetic data via foundation model apis 1: Images. *International Conference on Learning Representations (ICLR)*, 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Z., Wang, J., Dao, T., Zhou, T., Yuan, B., Song, Z., Shrivastava, A., Zhang, C., Tian, Y., Re, C., et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pp. 22137–22176. PMLR, 2023.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE Computer Society, 2023.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023.

Mattern, J., Jin, Z., Weggenmann, B., Schoelkopf, B., and Sachan, M. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4860–4873, 2022a.

Mattern, J., Weggenmann, B., and Kerschbaum, F. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 867–881, 2022b.

Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023.

MistralAI. Mixtral of experts. https://mistral.ai/news/mixtral-of-experts/, 2022.

OpenAI. ChatGPT. https://chat.openai.com, 2022.

OpenAI. Gpt-3.5 turbo fine-tuning and api updates, 2023a. URL https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023b.

OpenAI. What are tokens and how to count them?, 2023c. URL https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

Putta, P., Steele, A., and Ferrara, J. W. Differentially private conditional text generation for synthetic data production. 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V., McCoy, T., and Perlis, R. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921–e921, 2016.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tramèr, F., Kamath, G., and Carlini, N. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Utpala, S., Hooker, S., and Chen, P.-Y. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8442–8457, 2023.

Voytovich, L. and Greenberg, C. Natural language processing: practical applications in medicine and investigation of contextual autocomplete. In *Machine Learning in Clinical Neuroscience: Foundations and Applications*, pp. 207–214. Springer, 2022.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pp. 12208–12218. PMLR, 2021.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. Differentially private fine-tuning of language models. *International Conference on Learning Representations*, 2022.

Yu, D., Backurs, A., Gopi, S., Inan, H., Kulkarni, J., Lin, Z., Xie, C., Zhang, H., and Zhang, W. Training private and efficient language models with synthetic data from llms. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2023.

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., Sun, H., Levitan, D., and Sim, R. Synthetic text generation with differential privacy: A simple and practical recipe. *ACL*, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A. Privacy Analysis

We first introduce a related theorem from Balle & Wang (2018) in Thm. 1.

Theorem 1 (Analytic Gaussian Mechanism (Balle & Wang, 2018))**.** *Let $f : \mathbb{X} \to \mathbb{R}^d$ be a function with global L_2 sensitivity Δ. For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, the Gaussian output perturbation mechanism $M(x) = f(x) + Z$ with $Z \sim \mathcal{N}\left(0, \sigma^2 I\right)$ is $(\varepsilon, \delta) - DP$ if and only if*

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \leq \delta.$$

Next, we provide the privacy guarantee for Alg. 1 in Thm. 2

Theorem 2 (Privacy Guarantee for Alg. 1)**.** *Let Alg. 1 run T iterations, with noise multiplier σ (noise is added to each bin of the histogram), the DP mechanism satisfies (ε, δ)-DP if and only if*

$$\Phi\left(\frac{\sqrt{T}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{T}}\right) - e^{\varepsilon}\Phi\left(-\frac{\sqrt{T}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{T}}\right) \leq \delta.$$

Proof Sketch. The proof is very similar to the one in Lin et al. (2024). So we just describe the key steps at a high level. The L_2 sensitivity of the histogram created in each iteration of Alg. 1 is $\Delta = 1$, to which we add Gaussian noise of scale σ. Therefore T iterations of the algorithm can be seen as the adaptive composition of T Gaussian mechanisms with L_2 sensitivity 1 and noise scale σ. The privacy loss of the composition is equivalent to that of a single Gaussian mechanism with L_2 sensitivity 1 and noise scale $\sigma/\sqrt{T}$ according to the adaptive composition theorem of Gaussian mechanisms (Corollary 3. of (Dong et al., 2019)). Therefore the privacy gaurantee follows from Theorem 1. □

B. Additional Experimental Details

B.1. Datasets and Downstream Tasks.

Table 10: Dataset details.

Dataset	# Train	# Val	# Test	label 1	label 2
Yelp	1.9M	5000	5000	business category (10 classes)	review ratings (5 classes)
OpenReview (ICLR2023)	8396	2798	2798	review area (12 classes)	review recommendation (5 classes)
PubMed (2023/08/01-2023/08/07)	75316	14423	4453	next-word prediction	

We evaluate AUG-PE on there datasets:

- Yelp: Yelp data is a public benchmark providing reviews on businesses, and we used the preprocessed Yelp from (Yue et al., 2023). The number of train/val/test samples and label information in Tb. 10.
- OpenReview: For OpenReview ICLR2023 data, we crawl the meta-data for each review using the OpenReview Python library,[8] and concatenate the fields “summary_of_the_paper”, “strength_and_weaknesses” and “summary_of_the_review” as one sample in our dataset. We group the two attributes – review area and recommendation – together as a combination, and drop the training samples from combinations that contain fewer than 50 training samples. The number of samples after such preprocessing and label information is provided in Tb. 10. The number of samples for each class is provided in Tb. 11 and Tb. 12.
- PubMed: we use PubMed with abstracts of medical papers[9] crawled by Yu et al. (2023) from 2023/08/01 to 2023/08/07. The number of train/val/test samples are reported in Tb. 10.

For Yelp and OpenReview, we focus on conditional generation and use two attributes (i.e., labels) for each dataset: the review ratings (ranging from 1 star to 5 stars) and business category for Yelp data, and the review recommendation (ranging from “1: strong reject” to “8: accept, good paper”) and review area for OpenReview ICLR2023 data. We then use those labels for downstream classification tasks based on synthetic texts.

For PubMed, we focus on unconditional generation and use next-word prediction as downstream tasks. This is motivated by (Yu et al., 2023)

[8] https://github.com/openreview/openreview-py

[9] https://www.ncbi.nlm.nih.gov/

Table 11: Area label statistics of OpenReview.

Class Name	# Train Samples (Proportion)	# Test Samples (Proportion)
Deep Learning and Representational Learning	2479 (29.53%)	854 (30.52%)
Applications (e.g., speech processing, computer vision, NLP)	1100 (13.10%)	380 (13.58%)
Reinforcement Learning (e.g., decision and control, planning, hierarchical RL, robotics)	1016 (12.10%)	344 (12.29%)
Social Aspects of Machine Learning (e.g., AI safety, fairness, privacy, interpretability, human-AI interaction, ethics)	765 (9.11%)	248 (8.86%)
General Machine Learning	598 (7.12%)	177 (6.33%)
Theory (e.g., control theory, learning theory, algorithmic game theory)	458 (5.45%)	144 (5.15%)
Unsupervised and Self-supervised Learning	452 (5.38%)	135 (4.82%)
Machine Learning for Sciences (e.g., biology, physics, health sciences, social sciences, climate/sustainability)	440 (5.24%)	166 (5.93%)
Generative Models	390 (4.65%)	119 (4.25%)
Optimization (e.g., convex and non-convex optimization)	318 (3.79%)	96 (3.43%)
Probabilistic Methods (e.g., variational inference, causal inference, Gaussian processes)	230 (2.74%)	81 (2.89%)
Neuroscience and Cognitive Science (e.g., neural coding, brain-computer interfaces)	150 (1.79%)	54 (1.93%)

Table 12: Rating label statistics of OpenReview.

Class Name	# Train Samples (Proportion)	# Test Samples (Proportion)
Recommendation: 6: marginally above the acceptance threshold	2870 (34.18%)	896 (32.02%)
Recommendation: 5: marginally below the acceptance threshold	2144 (25.54%)	760 (27.16%)
Recommendation: 3: reject, not good enough	1703 (20.28%)	571 (20.41%)
Recommendation: 8: accept, good paper	1629 (19.40%)	554 (19.80%)
Recommendation: 1: strong reject	50 (0.60%)	17 (0.61%)

B.2. Implementation Details of AUG-PE.

B.2.1. MODEL AND HYPERPARAMETERS

We consider four LLMs as data generators in AUG-PE via API-access: GPT-2 (Radford et al., 2019), GPT-2-Medium, GPT-2-Large, and GPT-3.5 ("gpt-35-turbo" hosted on Microsft Azure[10]) (OpenAI, 2022). We provide the default hyper-parameter setup for GPT-3.5 in Tb. 13 and GPT-2 series models in Tb. 14.

The embedding model Φ in AUG-PE is instantiated by the sentence-transformer from HuggingFace. We use "stsbroberta-base-v2" for OpenReview and Yelp and "sentence-t5-base" for PubMed.

After generating the synthetic samples, we remove those with fewer than 100/50 tokens for OpenReview/PubMed. We noticed that samples with token lengths below those thresholds usually result from an unsuccessful API call for paper review/medical abstract generation (e.g. GPT-3.5 refuses to answer).

In terms of downstream models,

- For Yelp and OpenReview, we finetune the pre-trained RoBERTa-base model for all downstream text classification tasks. We set the max sequence length as 512, the batch size as 64, the learning rate as 3e-5, and the number of epochs as 5 for Yelp and 10 for OpenReview.

- For PubMed, we leverage pre-trained $\text{BERT}_{\text{Mini}}$ and $\text{BERT}_{\text{Small}}$ released by (Turc et al., 2019), which are lightweight to meet the inference time and computational cost requirements in many real-use cases. These models employ WordPiece tokenization and were trained on Wikipedia and BookCorpus using masked language modeling. During our downstream task fine-tuning, we implement a causal language modeling mask, restricting each token to attend only to its preceding tokens (Yu et al., 2023). We set the max sequence length as 512, batch size as 32, learning rate as 3e-4, the weight decay as 0.01. We finetune 20 epochs for $\text{BERT}_{\text{Mini}}$ and 10 for $\text{BERT}_{\text{Small}}$ epochs.

Table 13: Hyperparameters for GPT-3.5.

	N_{syn}	K	VARIATION_API	mask prob. p%	L	PE iteration	temperature	w2t_ratio	σ_{word}	min_word	max_token for RANDOM_API
Yelp	5k	3	fill-in-the-blanks (3-shot)	50%	1	20	1.4	1.2	40	25	128
OpenReview	2k	0	fill-in-the-blanks (1-shot)	50%	4	10	1.2	5	60	25	1000
PubMed	2k	0	fill-in-the-blanks (0-shot)	50%	4	10	1.2	5	60	25	1000

[10] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models

Table 14: Hyperparameters for GPT-2, GPT-2-Medium, and GPT-2-Large.

Model	N_{syn}	K	VARIATION_API	L	PE iteration T	temperature	$\max\_\text{token}$
Yelp	5k, 10k, 100k	0	paraphrasing (zero-shot)	7	20	1.4	64
OpenReview	2k, 3k, 5k	0	paraphrasing (zero-shot)	7	10	1.2	448
PubMed	2k, 3k, 5k	0	paraphrasing (zero-shot)	7	10	1.0	448

Table 15: Prompts as RANDOM_API for GPT-3.5.

Speaker	**Yelp**	**OpenReview**	**PubMed**
System	You are required to write an example of review based on the provided Business Category and Review Stars that fall within the range of 1.0-5.0.	Given the area and final decision of a research paper, you are required to provide an example of the review consisting of the following content: 1. briefly summarizing the paper in 3-5 sentences; 2. listing the strengths and weaknesses of the paper in details; 3. briefly summarizing the review in 3-5 sentences.	Please act as a sentence generator for the medical domain. Generated sentences should mimic the style of PubMed journal articles, using a variety of sentence structures.
User	Business Category: {label_1} | Review Stars: {label_2} with keyword {subcategory}	Area: {label_1} | Recommendation: {label_2}	Suppose that you are a {writer}. Please write an abstract for a medical research paper:

B.2.2. API Prompt Designs

In terms of RANDOM_API,

- For Yelp data, we generate 100 subcategories under each business category via ChatGPT and use them as keywords in the prompts.
- For OpenReview data, we do not generate subcategories, as the review area label (e.g., "*Social Aspects of Machine Learning (eg, AI safety, fairness, privacy, interpretability, human-AI interaction, ethics)*") already provides detailed information about the area. Instead, we generate a list of writers with their corresponing tones via ChatGPT (e.g., "*Postdoctoral Researcher: Advanced and knowledgeable insights", "AI Policy Maker: Concerned with regulatory and policy implications", "Robotics Engineer: Focus on practical applications in robotics"*) and use them as keywords in the prompt.
- For PubMed data, we also generate a list of writers for medical abstracts via ChatGPT, such as "*Clinical Researcher, Principal Investigator, Biomedical Engineer*", etc., and use them as keywords in the prompt.

We provide the prompts of RANDOM_API for all datasets in Tb. 15 for GPT-3.5 and Tb. 16 for other LLMs.

In terms of VARIATION_API, (1) for GPT-3.5, we utilize fill-in-the-blanks with adaptive text lengths, providing few-shot demonstrations. To obtain {masked_input} used for fill-in-the-blanks, we calculate the tokens for {input} based on GPT-3.5 tokenizer[11], mask $p\%$ of them as blanks "_", and decode them back to the text. (2) In contrast, for GPT-2-series models, we opt for zero-shot paraphrasing with fixed $\max\_\text{token}$ as VARIATION_API. This choice is based on our observation that GPT-2-series models do not follow the instructions of fill-in-the-blanks and adaptive text lengths well, as they are only pretrained on next-word-prediction tasks without further instruction tuning or reinforcement learning from human feedback (RLHF) (Lambert et al., 2022) for blank filling tasks. Moreover, GPT-2-series models do not gain much from few-shot demonstrations for paraphrasing, possibly due to their inferior instruction-following and in-context learning capabilities compared to GPT-3.5.

We provide the prompts of VARIATION_API for GPT-2-series models in Tb. 17 and for GPT-3.5 in Tb. 18, Tb. 19 and Tb. 20.

[11] https://github.com/openai/tiktoken

Table 16: Prompts as RANDOM_API for GPT-2-series models.

Yelp	**OpenReview**	**PubMed**
Business Category: {label_1} | Review Stars: {label_2} with keyword {subcategory}	Suppose that you are a {writer}. Write a paper review based on Area: {label_1} | Recommendation: {label_2}	Using a variety of sentence structures, write an abstract for a medical research paper:

Table 17: Prompts as VARIATION_API for GPT-2-series models on Yelp and OpenReview.

Datast	**Prompt**
Yelp	Based on "Business Category: {label_1} | Review Stars: {label_2}", please rephrase the following sentences {in a selected_tone}: {input}
OpenReview	Based on "Area: {label_1} | Recommendation: {label_2}", please rephrase the following sentences {in a selected_tone}: {input}
PubMed	Please rephrase the following sentences {in a selected_tone} as an abstract for medical research paper: {input}

Table 18: Prompts as VARIATION_API for GPT-3.5 on Yelp.

Speaker	**Prompt**
System	You are a helpful, pattern-following assistant.
User	Based on the Business Category and Review Stars, you are required to fill in the blanks in the Input sentences. If there are no blanks, you are required to output the original Input sentences. Business Category: Restaurants | Review Stars: 2.0 Input: _ that great , terrible _ rolls and fish _ smelling _ _. Fill-in-Blanks and your answer MUST be exactly 10 words: Not that great, terrible egg rolls and fishy smelling shrimp. Business Category: Beauty & Spas | Review Stars: 5.0 Input: Very clean! Staff are super friendly!! Fill-in-Blanks and your answer MUST be exactly 6 words: Very clean! Staff are super friendly!! Business Category: Shopping | Review Stars: 3.0 Input: I _ in _ and stopped in for a _. I was _ surprised. Good _, nice price. Fill-in-Blanks and your answer MUST be exactly 19 words: I was in a rush and stopped in for a mani-pedi. I was pleasantly surprised. Good service, nice price. Business Category: {label_1} | Review Stars: {label_2} Input: {masked_input} Fill-in-Blanks and your answer MUST be exactly {targeted_word} words:

B.2.3. DIFFERENTIAL PRIVACY.

Following Yue et al. (2023), we set $\delta = \frac{1}{N_{\text{priv}} \cdot \log(N_{\text{priv}})}$ for (ϵ, δ)-DP. As different datasets have different sizes of private training data, they require different δ. We run 10 PE iterations under DP on all datasets. To achieve $\epsilon = \{1, 2, 4, \infty\}$, we use noise multiplier $\sigma = \{15.34, 8.03, 4.24, 0\}$ for Yelp; $\sigma = \{11.60, 6.22, 3.38, 0\}$ for OpenReview; $\sigma = \{13.26, 7.01, 3.75, 0\}$ for PubMed.

Table 19: Prompts as VARIATION_API for GPT-3.5 on OpenReview.

Speaker	Prompt
System	You are an AI assistant that helps people find information.
User	Based on the area and final recommendation of a research paper, you are required to fill in the blanks for the input sentences {in a selected_tone}. If there is no blanks, please output the original input sentences. Area: Applications (eg, speech processing, computer vision, NLP) | Recommendation: 3: reject, not good enough Input: __ proposes an__ method_ ROI detection__arial_f_ without attention_. The_ map can_ used____ for__ and____ show_ improvements on different medical__._Strength__ \n–The idea using__actual images_ sali__ generation_ interesting.\n\n_The improvement____aks is significant. \n\nWeak____The___ and_____ experiments are needed_ such as__f___the_ method_ interesting_ but_ novelty_ limited Fill-in-Blanks and your answer MUST be exactly 85 words: This paper proposes an attention generation method for ROI detection by adversarial counterfactual without attention label. The attention map can be used to highlight useful information for disease classification and detection. The experiments show its improvements on different medical imaging tasks. \nStrengths: \n–The idea using counterfactual images for saliency map generation is interesting.\n\n–The improvement for medical imaging taks is significant. \n\nWeaknesses:\n\n–The novelty is simple and limited. \n\n–More experiments are needed, such as existing counterfactual generation.\nthe proposed method is interesting, but the novelty is limited. Area: {label_1} | Recommendation: {label_2} Input: {masked_input} Fill-in-Blanks and your answer MUST be exactly {targeted_word} words:

Table 20: Prompts as VARIATION_API for GPT-3.5 on PubMed.

Speaker	Prompt
System	Please act as a sentence generator for the medical domain. Generated sentences should mimic the style of PubMed journal articles, using a variety of sentence structures.
User	You are required to fill in the blanks with more details for the input medical abstract {in a selected_tone}. If there is no blanks, please output the original medical abstract. Please fill in the blanks in the following sentences to write an abstract of a medical research paper: {masked_input} and your answer MUST be exactly {targeted_word} words.

B.3. Implementation Details of Baselines.

For DP-FT-GENERATOR, we finetune the GPT-2-series models following the hyperparameters setup in Table 8 of (Yue et al., 2023).

For DP-FT-DOWNSTREAM, we report the hyperparameters for OpenReview and Yelp in Tb. 22, and PubMed in Tb. 21. For a target ϵ, a noise multiplier is set as the smallest value such that DP-SGD can run the target number of steps.

Table 21: Hyperparameters for DP-FT-DOWNSTREAM on PubMed.

	$\text{BERT}_{\text{Tiny}}$, $\text{BERT}_{\text{Mini}}$, $\text{BERT}_{\text{Small}}$ for PubMed		LLaMA-2-7B for PubMed	
	downstream (non-pri.)	downstream (pri.)	downstream (non-pri.)	downstream (pri.)
Epoch	[5, 10, 30]	[10, 30, 50, 100]	10	10
Batch size	[32, 64]	[1024, 2048, 4096]	128	128
Clipping norm	-	[0.1, 0.5, 1, 3, 5]	-	1
Learning rate	$[3 \times 10^{-5}, \{1, 3\} \times 10^{-4}]$	$[3 \times 10^{-4}, \{1, 3\} \times 10^{-3}]$	1×10^{-3}	1×10^{-3}

B.4. Metrics.

Here we provide more details about the metrics regarding embedding distribution distance. We use sentence-transformer "stsb-roberta-base-v2" from HuggingFace[12] to embed the real and synthetic datasets, and use seven evaluation metrics to measure embedding distribution distance: 1) Fréchet Inception Distance (*FID*) evaluates the feature-wise mean and

[12] https://huggingface.co/models

Table 22: Hyperparameters for DP-FT-DOWNSTREAM on Yelp and OpenReview.

	RoBERTa-base for Yelp		RoBERTa-base for OpenReview	
	downstream (non-pri.)	downstream (pri.)	downstream (non-pri.)	downstream (pri.)
Epoch	[1,10]	[1,10]	10	10
Batch size	[128, 1024]	[128, 1024]	8	128
Clipping norm	-	1	-	1
Learning rate	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}

covariance matrices of the embedding vectors and then computes the Fréchet distance between these two groups (Heusel et al., 2017) ; 2) *Precision* estimates the average sample quality; 3) *Recall* assesses the breadth of the sample distribution; 4) *F1* score is the harmonic mean of Precision and Recall, serving as a balance of the two (Kynkäänniemi et al., 2019); 5) MAUVE evaluates the distributional distance of the synthetic and real data via divergence frontiers (Pillutla et al., 2021); 6) *KL div.* measures the distance of embedding distributions based on KL divergence; 7) *TV div.* quantifies the distance based on Total Variation divergence (Chung et al., 1989).

For downstream classification accuracy, we train downstream models **three times** and report the average accuracy. For each metric associated with embedding distribution distance (except FID for which we use the whole dataset), we randomly draw 5000 samples (for efficiency) from the private dataset and the synthetic dataset respectively, to calculate the distance. We then report the averaged results based on **five** independent draws.

C. Additional Experimental Results

C.1. Robustness Against Membership Inference Attacks

In this work, we focus on DP, a type of widely accepted privacy guarantee with profound theoretic backup which provides an upper bound for empirical membership privacy attacks. To better understanding empirical risk, we perform state-of-the-art membership inference attacks (MIAs) (Shokri et al., 2017) in the text domain.

We perform MIAs against the finetuned downstream models (which are fine-tuned on synthetic data for AUG-PE/ DP-FT-GENERATOR; on real private data for DP-FT-DOWNSTREAM). We randomly sample 4000 PubMed real private samples as members and 4000 PubMed test samples as non-members for evaluation. We report AUC (Area Under the Curve) to evaluate the risks of MIAs. We consider three types of MIAs: (1) PPL thresholds perplexity to predict membership (Carlini et al., 2021). (2) REFER computes the ratio of the log perplexity of the tested model against a reference model (Carlini et al., 2021). (3) LIRA uses the ratio of likelihood instead of log-perplexity (Carlini et al., 2022). LiRA assumes the availability of high-quality data distributed similarly to the training set, which was thought to be impractical (Tramèr et al., 2022). Therefore, we follow (Mattern et al., 2023) to use the pre-trained model as a reference.

The results in Tb. 3 show that AUG-PE exhibits lower MIA AUC scores compared to both DP-FT-GENERATOR and DP-FT-DOWNSTREAM models. This indicates a higher robustness to empirical privacy attacks, potentially due to the synthetic nature of the data used for downstream model finetuning, which inherently reduces the risk of overfitting to real private data.

C.2. Convergence of Text Length Distribution

As shown in Fig. 7, Fig. 8 and Fig. 9[13], we see that over the PE iterations, the text length distribution of synthetic samples produced from GPT-3.5 through our AUG-PE converges, as it becomes closer to the distribution of the original data. This showcases the effectiveness of our adaptive text length mechanism. We note that there is a noticeable peak near 30 tokens for our synthetic texts on Yelp, which is attributed to the $\min\_\text{word}$ used in the VARIATION_API prompt to avoid generating blank outputs.

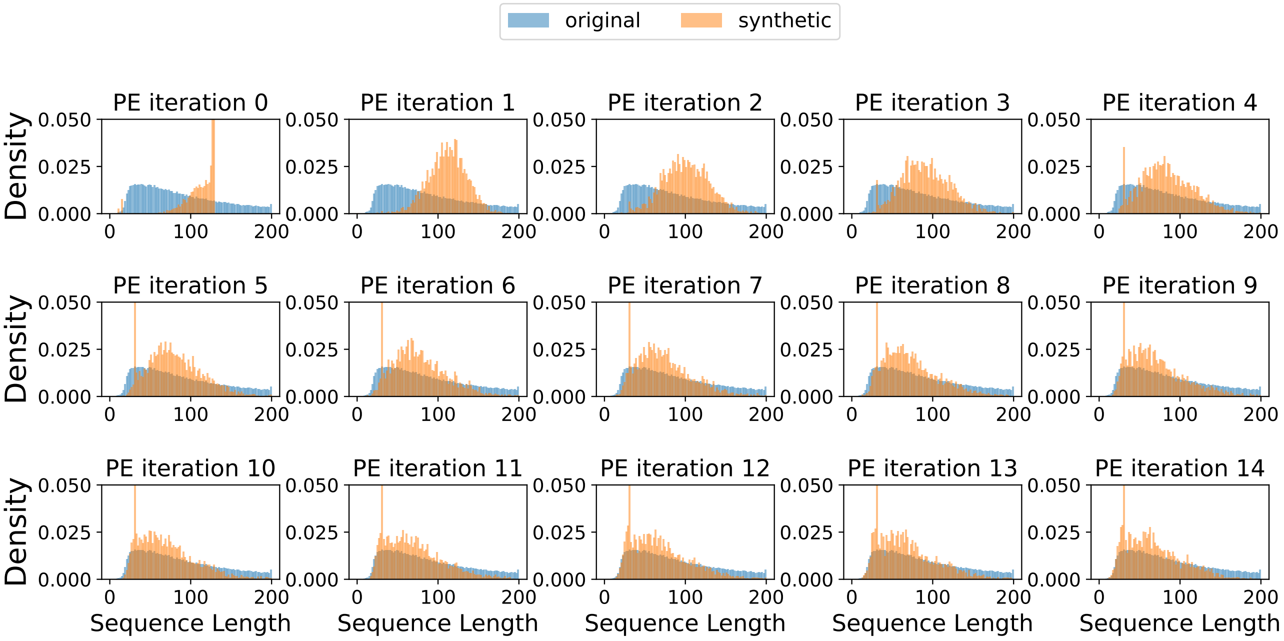

Figure 7: Convergence of text length distribution over AUG-PE iterations on Yelp synthetic text generated from GPT-3.5.

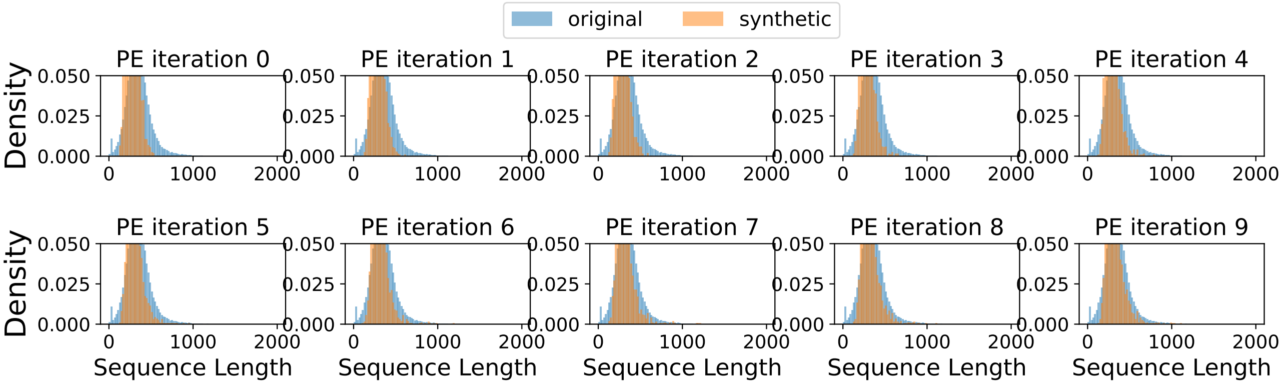

Figure 8: Convergence of text length distribution over AUG-PE iterations on PubMed synthetic text generated from GPT-3.5.

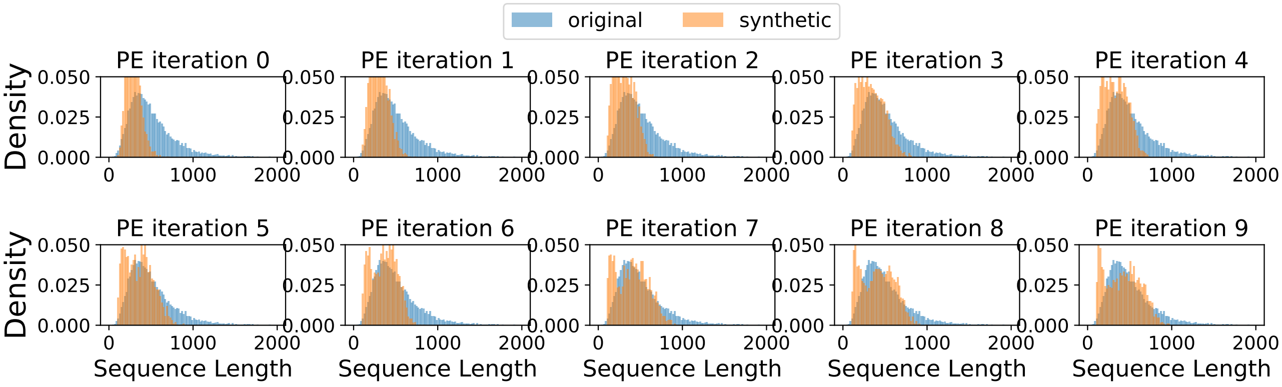

Figure 9: Convergence of text length distribution over AUG-PE iterations on OpenReview synthetic text generated from GPT-3.5.

Table 23: GPU hours on one 32G NVIDIA V100 for AUG-PE DP-FT-GENERATOR on Yelp under $\epsilon = 1$. AUG-PE is more efficient with fewer total GPU hours.

		DP-SGD finetune	Generation		
			5k samples	10k samples	100k samples
DP-FT-GENERATOR	GPT2	456.71	0.22	0.45	4.47
	GPT2-Medium	709.50	0.25	0.50	5.03
	GPT2-large	1764.42	0.35	0.70	6.96
AUG-PE ($L = 2$)	GPT2	/	1.76	2.48	13.35
	GPT2-Medium	/	2.30	2.89	18.68
	GPT2-large	/	2.68	3.83	26.98
AUG-PE ($L = 7$)	GPT2	/	6.04	9.07	66.66
	GPT2-Medium	/	6.94	11.55	91.07
	GPT2-large	/	9.62	16.77	139.35

C.3. Efficiency in Terms of GPU Hours

In Tb. 23, we provide a detailed breakdown of the GPU hours shown in Fig. 3. We consider the process of generating DP synthetic data given a private dataset. DP-FT-GENERATOR (Yue et al., 2023) requires two steps: (1) finetuning a pretrained data generator with DP-SGD, and (2) generating samples from the finetuned data generator, whereas AUG-PE requires only one step (Alg. 1). In Tb. 23, we list the GPU hours of each step of each method. For Yue et al. (2023), we use the hyper-parameters in their Table 8.

We can see that the majority of the time spent by DP-FT-GENERATOR is the DP-fine-tuning stage, which is already much more costly than the total cost of AUG-PE. This results from two factors: (1) Training is costly due to the backpropagation, especially for large models; (2) DP-SGD requires per-sample gradients, which further increases the memory and computation cost. In contrast, AUG-PE only requires model inference and does not require model training, and is thus more efficient.

It is also worth noting that once the model is DP finetuned, DP-FT-GENERATOR can efficiently generate many samples with only model inference. It is illustrated by the small GPU hours in the "Generation" step of DP-FT-GENERATOR. In contrast, in AUG-PE, the required GPU hour is positively correlated with the number of samples. Therefore, DP-FT-GENERATOR can become more efficient than AUG-PE when the number of generated samples is large enough. However, the original PE paper (Lin et al., 2024) proposed an efficient way to generate more DP samples after PE is done, by passing the generated samples through VARIATION_API. In the context of text generation with LLMs, this approach is expected to have a similar overhead as generating more samples from the DP-finetuned generator in DP-FT-GENERATOR. We defer the study of this approach to future work.

C.4. Comparison Between AUG-PE and PE

We compare AUG-PE against PE when using GPT-3.5 as the generator on three datasets. The results in Tb. 24 show that AUG-PE is always better than PE on PubMed for GPT-3.5. Moreover, AUG-PE is better for OpenReview Rating classification task and Yelp Rating classification task. As AUG-PE supports PE as a special case by changing the hyperparameters of L and K, the practitioner can adjust those hyperparameters for a specific downstream task and find the best settings to generate synthetic data.

C.5. Ablation Study on Variation API Prompt Design

The results in Tb. 25 show that fill-in-the-blanks prompt (with few-shot demonstrations) yields better results for GPT-3.5. For GPT-2, paraphrasing can be an effective strategy. Although fill-in-blanks leads to high accuracy on Yelp Category classification task, we find that the generated texts have many unfilled blanks "__" upon inspection.

[13] For OpenReview in Fig. 9, we use a temperature of 1.4.

Table 24: Comparision between AUG-PE and PE when using GPT-3.5 as generator on three datasets.

Data Type (Size)	Method	$\epsilon = \infty$		$\epsilon = 4$		$\epsilon = 2$		$\epsilon = 1$	
Yelp		Rating	Category	Rating	Category	Rating	Category	Rating	Category
Synthetic (5000)	PE ← AUG-PE ($k = 3, L = 1$)	67.9	**74.7**	67.1	**74.6**	67.2	**74.6**	67.6	**74.7**
Synthetic (5000)	AUG-PE ($k = 0, L = 4$)	**68.4**	74.1	**68.1**	74.0	**67.8**	74.3	**67.9**	74.0
OpenReview		Area	Rating	Area	Rating	Area	Rating	Area	Rating
Synthetic (2000)	PE ← AUG-PE ($k = 3, L = 1$)	43.6	42.4	**43.6**	43.5	**44.6**	43.7	**42.0**	42.9
Synthetic (2000)	AUG-PE ($k = 0, L = 4$)	**45.4**	**43.5**	43.5	**44.6**	42.8	**44.5**	41.9	**43.1**
PubMed		$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$	$\text{BERT}_{\text{Mini}}$	$\text{BERT}_{\text{Small}}$
Synthetic (2000)	PE ← AUG-PE ($k = 3, L = 1$)	29.7	31.8	29.6	31.8	29.7	31.9	29.8	31.9
Synthetic (2000)	AUG-PE ($k = 0, L = 4$)	**30.4**	**32.7**	**30.3**	**32.5**	**30.2**	**32.5**	**30.1**	**32.4**

Table 25: Evaluation on Variation API designs for GPT-2 and GPT-3.5 on Yelp. Fill-in-the-blanks is prefered for GPT-3.5.

Variation API prompt	GPT-2		GPT-3.5	
	Rating	Category	Rating	Category
paraphrasing	67.5	74.8	67.5	74.3
paraphrasing w/ few-shot demos	**67.8**	73.6	65.7	74.2
fill-in-the-blanks	66.3	**74.6**	**67.9**	74.6
fill-in-the-blanks w/ few-shot demos	67.6	74.8	**67.9**	**74.7**

C.6. Leveraging Open-source LLMs as Generator for AUG-PE

We use opensource LLMs from Huggingface as data generators in AUG-PE. We find that LLaMA-2-7B does not follow the fill-in-the-blank prompts well and often leaves blanks ("__") in the generated texts. Also, it struggles to adhere to the word prompt and the length of synthetic sequences exhibits a large gap from the targeted word specified in the prompt. It might be because they are not explicitly instruction/RLHF-tuned for those blank-filling and word count tasks, and have inferior instruction-following and in-context learning capabilities compared to GPT-3.5. Therefore, we turn to use the same hyperparameter setup as GPT-2-series models for those open-source LLMs. The results in Tb. 2 show that GPT-3.5 outperforms most of the models on PubMed tasks and OpenReview Rating classification task by a large margin. For OpenReview Area task, Mixtral-8x7B-v0.1 is better than GPT-3.5, demonstrating the competitive generation power of Mixtral-8x7B-v0.1 for academic reviews in machine learning domains.

C.7. Effect of Rank-based Sampling

Table 26: Comparing rank-based sampling against probability-based random sampling for AUG-PE with GPT-2-series models on three datasets.

Data Type (Synthetic Data Size)	Data Generator	Rank	Prob	Rank	Prob
		Rating		Category	
Yelp (5000)	GPT-2	**67.5**	66.7	**74.8**	74.7
	GPT-2 Medium	67.5	**67.7**	**74.9**	74.6
	GPT-2 Large	**67.5**	67.1	**74.5**	74.4
		Area		Rating	
OpenReview (2000)	GPT-2	**42.4**	39.8	32.1	32.1
	GPT-2 Medium	**41.0**	37.1	**32.3**	32.0
	GPT-2 Large	**42.1**	40.1	**32.1**	32.0
		$\text{BERT}_{\text{Mini}}$		$\text{BERT}_{\text{Small}}$	
PubMed (2000)	GPT-2	**24.5**	23.4	**26.7**	25.4
	GPT-2 Medium	**25.5**	23.9	**27.7**	25.9
	GPT-2 Large	**25.7**	24.1	**28.0**	26.0

We compare our proposed rank-based sampling (Line 19) against probability-based random sampling in the original PE

(Line 15) across GPT-2, GPT-2-Medium and GPT-2-Large on three datasets. The results in Tb. 26 indicate that our proposed rank-based sampling (Line 19) consistently outperforms probability-based random sampling in the original PE (Line 15), due to the elimination of sample redundancy inherent in random sampling, as rank-based sampling exclusively selects the top N_{syn} samples.

C.8. Effect of Iteration T on DP Utility

Tb. 6 presents the results on the non-DP setting, serving as an ablation study to underscore the role of private data in AUG-PE. In the DP setting, given a fixed privacy budget, a larger T requires more noise, which may compromise the utility of the DP histogram. On the other hand, a larger T allows for more iterations of sample improvement. We want to study the joint effect of these two factors.

We conducted experiments comparing the utility of the algorithm at $t = 1$ and $t = 10$ on three datasets under $\epsilon = 4, 2, 1$. The results in Tbs. 27 to 29 show that $t = 10$ consistently yields better utility than $t = 1$ across all three privacy budget levels, underscoring the effectiveness of AUG-PE's iterative improvement mechanism. This finding suggests that, despite the increased noise, the algorithm can robustly preserve useful statistical properties in the DP histogram and generate high-quality texts under $t = 10$.

Table 27: Effect of AUG-PE iteration t on the DP utility under $\epsilon = 4$.

	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	BERT$_{\text{Mini}}$	BERT$_{\text{Small}}$
$t = 1$	64.7	73.5	36.5	42.1	29.9	32.3
$t = 10$	67.8	74.6	43.5	44.6	30.3	32.5

Table 28: Effect of AUG-PE iteration t on the DP utility under $\epsilon = 2$.

	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	BERT$_{\text{Mini}}$	BERT$_{\text{Small}}$
$t = 1$	63.9	73.6	37.2	42.0	29.9	32.3
$t = 10$	67.4	74.3	42.8	44.5	30.2	32.5

Table 29: Effect of AUG-PE iteration t on the DP utility under $\epsilon = 1$.

	Yelp		OpenReview		PubMed	
	Rating	Category	Area	Rating	BERT$_{\text{Mini}}$	BERT$_{\text{Small}}$
$t = 1$	63.8	73.1	37.4	41.7	30.0	32.3
$t = 10$	66.8	74.7	41.9	43.1	30.1	32.4

C.9. Embedding Distribution Distance Between Real and Synthetic data

We report the results of embedding distribution distance between real and synthetic data on Yelp in Fig. 10, and on PubMed in Fig. 11. When using the same base model GPT-2 for a fair comparison, we observe that under DP and non-DP settings, AUG-PE can obtain similar and even lower embedding distribution distances between real and synthetic samples for certain metrics compared to fine-tuning. For example, on Yelp dataset, under DP, AUG-PE yields better FID, precision, recall, F1 than DP-FT-GENERATOR and achieves comparable MAUVE scores. On PubMed dataset, under DP, AUG-PE yields better FID, MAUVE scores, KL divergence, and TV divergence than DP-FT-GENERATOR. These findings highlight the promise of employing the API-only method for DP synthetic text generation.

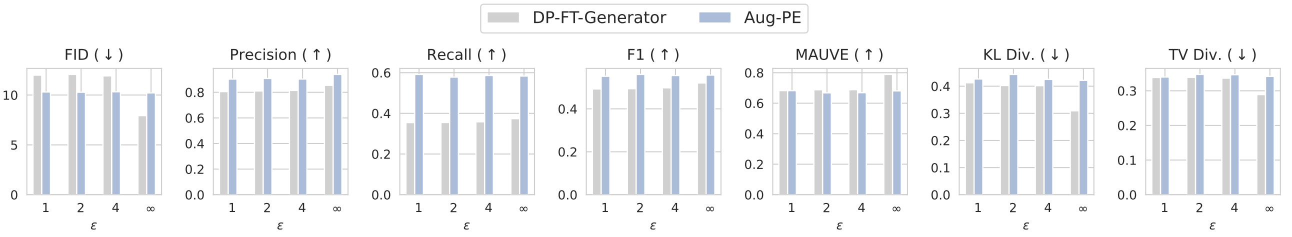

Figure 10: Evaluation on distribution distances between Yelp real data and GPT-2 generated 10k DP synthetic samples.

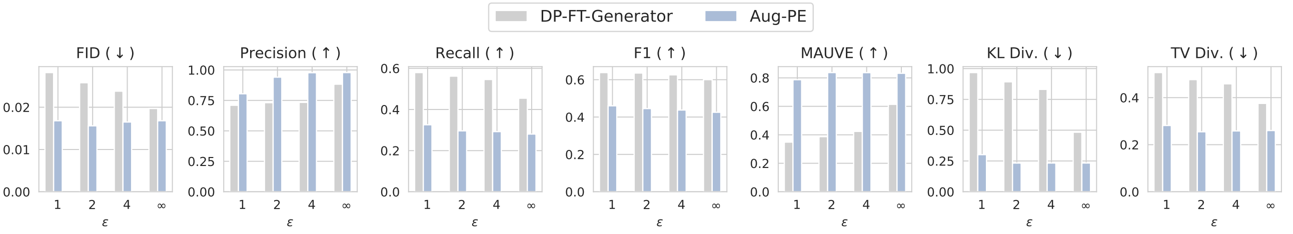

Figure 11: Evaluation on distribution distances between PubMed real data and GPT-2 generated 2k DP synthetic samples.

C.10. Downstream Task Utility Under Various Synthetic Data Size

C.10.1. Utility on Yelp

We report the full results of downstream accuracy on Yelp in Tb. 30. We find that (1) when using the same base model for a fair comparison, we see that under DP settings, AUG-PE demonstrates competitive (or even better) utility on downstream classification tasks compared to fine-tuning. The scores are also close to that of the downstream algorithms trained on the real data under DP directly, demonstrating the promise of DP synthetic text as a tool for DP machine learning. (2) For large models like GPT-2-Large and GPT-2-Medium, more synthetic samples (e.g., 100k) from AUG-PE can enhance downstream utility. However, for GPT-2, sometimes 10k synthetic samples can lead to better downstream utility than 100k samples, which might be due to the low-quality data generated from the small model that hurts the performance.

C.10.2. Utility on OpenReview

We report the downstream accuracy on OpenReview in Tb. 31. The key observations are: (1) Under DP when using the same GPT-2/GPT-2-Medium/GPT-2-Large as the base model, AUG-PE achieve similar classification accuracy and classification accuracy compared with DP-FT-GENERATOR. This again demonstrates that AUG-PE is a promising alternative to DP fine-tuning. (2) More synthetic samples lead to better area classification accuracy for the three GPT-2-series models, indicating that AUG-PE scales well with the synthetic sample size. Note that both AUG-PE and DP-FT-GENERATOR do not perform well on review rating classification tasks across different data sizes, which shows the inherent limitation of GPT-2-series models – they may struggle to generate academic texts with correct sentiments. (3) AUG-PE with GPT-3.5 achieves better utility than AUG-PE with GPT-2-Large on both tasks with or without DP. This suggests that AUG-PE benefits from larger and more powerful LLMs. We expect that as the capability of LLMs quickly evolves, AUG-PE can be even more promising in the future. (4) However, there is still a gap between the results of AUG-PE under non-DP setting $\epsilon = \infty$ and the results on the original data. This suggests that even in the non-DP setting, AUG-PE is still not able to recover the distribution of the real data. This gap is unavoidable in the DP setting. We hypothesize that better hyper-parameter tunings (e.g., the variation degree) could lower the gap. We leave a more careful investigation of this issue to future work.

C.10.3. Utility on PubMed

We report the next-word prediction accuracy on OpenReview of downstream model $\text{BERT}_{\text{Mini}}$ in Tb. 32 and $\text{BERT}_{\text{Small}}$ in Tb. 33 We find that (1) under the same GPT-2-series model as generator, AUG-PE underperforms DP-FT-GENERATOR on PubMed. This is expected because AUG-PE relies on the knowledge within LLMs to generate high-quality texts without domain-specific finetuning, while GPT-2-series models might have limited exposure to biomedical literature (Radford et al., 2019). (2) With powerful LLMs like GPT-3.5, AUG-PE can outperform DP-FT-GENERATOR under DP. (3) Additionally,

Table 30: Classification accuracy of downstream RoBERTa-base model under $\epsilon = \infty, 4, 2, 1$ on Yelp for two downstream tasks: review rating and business category classification. (i) Compared to DP-FT-Generator, in some cases, downstream accuracy of Aug-PE is higher (↑) under the same synthetic data size and the same GPT-2-series data generator. Leveraging the inherent knowledge within stronger LLM, GPT-3.5, Aug-PE can achieve higher accuracy. (ii) Compared to traditional method DP-FT-Downstream, Aug-PE can also obtain higher accuracy under DP with the same synthetic data size.

Data Type (Size)	Method	Data Generator	$\epsilon = \infty$		$\epsilon = 4$		$\epsilon = 2$		$\epsilon = 1$	
			Rating	Category	Rating	Category	Rating	Category	Rating	Category
Original (1,939,290)	DP-FT-Downstream	-	76.0	81.6	67.5	72.8	67.2	72.0	66.8	71.8
Original (100,000)	DP-FT-Downstream	-	72.7	75.5	65.0	71.2	64.1	70.0	62.9	68.7
Original (10,000)	DP-FT-Downstream	-	70.9	76.2	44.8	61.8	44.8	61.8	44.8	61.8
Original (5,000)	DP-FT-Downstream	-	70.5	75.1	44.8	61.8	44.8	61.8	44.8	61.8
Synthetic (5000)	DP-FT-Generator	GPT-2	70.3	75.9	68.2	74.1	67.2	73.1	66.4	73.9
Synthetic (10000)	DP-FT-Generator	GPT-2	71.1	75.8	68.2	73.0	67.7	73.2	66.7	73.7
Synthetic (100000)	DP-FT-Generator	GPT-2	71.0	75.6	66.8	72.6	67.0	72.3	65.5	71.8
Synthetic (5000)	Aug-PE	GPT-2	67.5	74.8	66.4	74.9 ↑	67.1	74.7 ↑	66.9 ↑	74.4 ↑
Synthetic (10000)	Aug-PE	GPT-2	67.2	75.1	66.6	75.3 ↑	66.2	74.9 ↑	66.0	74.6 ↑
Synthetic (100000)	Aug-PE	GPT-2	67.1	76.0 ↑	66.3	75.1 ↑	66.1	75.0 ↑	65.7 ↑	74.5 ↑
Synthetic (5000)	DP-FT-Generator	GPT-2-Medium	70.0	75.0	69.1	74.6	67.8	74.3	67.4	74.1
Synthetic (10000)	DP-FT-Generator	GPT-2-Medium	70.7	75.6	68.8	74.4	68.2	73.8	67.5	73.9
Synthetic (100000)	DP-FT-Generator	GPT-2-Medium	71.9	76.3	68.1	73.9	67.8	74.3	67.9	73.3
Synthetic (5000)	Aug-PE	GPT-2-Medium	67.5	74.9	66.8	74.6	67.8	74.7 ↑	67.4	74.6 ↑
Synthetic (10000)	Aug-PE	GPT-2-Medium	67.5	74.9	67.4	74.9 ↑	67.6	75.1 ↑	67.1	74.7 ↑
Synthetic (100000)	Aug-PE	GPT-2-Medium	68.2	75.8	67.4	75.5 ↑	66.6	75.3 ↑	66.2	74.7 ↑
Synthetic (5000)	DP-FT-Generator	GPT-2-Large	70.4	75.4	68.7	74.2	69.8	75.1	68.7	74.6
Synthetic (10000)	DP-FT-Generator	GPT-2-Large	70.7	74.3	69.2	74.9	69.7	75.2	68.9	74.6
Synthetic (100000)	DP-FT-Generator	GPT-2-Large	71.8	74.1	69.5	74.5	68.7	74.5	69.6	74.4
Synthetic (5000)	Aug-PE	GPT-2-Large	67.5	74.5	67.3	74.4 ↑	65.8	74.1	66.6	75.0 ↑
Synthetic (10000)	Aug-PE	GPT-2-Large	67.1	74.7 ↑	67.1	74.9	66.6	74.7	67.0	74.4
Synthetic (100000)	Aug-PE	GPT-2-Large	67.3	75.8 ↑	67.6	75.7 ↑	66.8	75.4 ↑	66.0	75.3 ↑
Synthetic (5000)	Aug-PE	GPT-3.5	68.4	74.1	68.1	74.0	67.8	74.3	67.9	74.0

more synthetic samples lead to better downstream classification accuracy for the three GPT-2-series models on PubMed.

C.11. Comparision to Text-to-Text Privatization Approaches

This is an active line of research on text-to-text privatization techniques for generating differentially private text. We do not directly compare these methods in our main paper due to the key distinctions in privacy definitions:

1. **Different privacy definitions**. Our method adopts the standard (ϵ, δ)-DP defined over neighboring datasets. This contrasts with
   (a) Word-level metric DP (Feyisetan et al., 2020; Carvalho et al., 2023): a specific metric for measuring word distance needs to be written in the privacy notation, and privacy guarantee is defined over neighboring words;
   (b) Local DP (Mattern et al., 2022b; Utpala et al., 2023): privacy guarantee is defined over neighboring samples.
2. **Poor privacy-utility trade-off in existing text-to-text anonymization methods**: While innovative, Feyisetan et al. (2020); Carvalho et al. (2023); Mattern et al. (2022b); Utpala et al. (2023) encounter challenges in achieving a good privacy-utility tradeoff under practical privacy budgets.
3. **Absence of privacy budgets**: The absence of detailed reporting on exact privacy budgets in (Utpala et al., 2023) hinders direct comparisons with our work.

A qualitative comparison between text-to-text privatization methods and our method is shown in Tb. 34.

Next, we compare Aug-PE with the text-to-text privatization frameworks in detail: word-level metric-DP frameworks (Feyisetan et al., 2020; Carvalho et al., 2023) and sample-level local-DP frameworks (Mattern et al., 2022b; Utpala et al., 2023).

Table 31: Classification accuracy of downstream RoBERTa-base model under $\epsilon = \infty, 4, 2, 1$ on OpenReview for two downstream tasks: review area and rating classification. (i) Compared to DP-FT-GENERATOR, in some cases, downstream accuracy of AUG-PE is higher (↑) under the same synthetic data size and the same GPT-2-series data generator. Leveraging the inherent knowledge within stronger LLM, GPT-3.5, AUG-PE can achieve higher accuracy. (ii) Compared to traditional method DP-FT-DOWNSTREAM, AUG-PE can also obtain higher accuracy under DP with the same synthetic data size.

Data Type (Size)	Method	Data Generator	$\epsilon = \infty$		$\epsilon = 4$		$\epsilon = 2$		$\epsilon = 1$	
			Area	Rating	Area	Rating	Area	Rating	Area	Rating
Original (8396)	DP-FT-DOWNSTREAM	-	65.2	50.9	30.5	32.0	30.5	32.0	30.5	32.0
Original (2000)	DP-FT-DOWNSTREAM	-	55.3	47.8	30.5	32.0	30.4	25.5	6.3	19.8
Synthetic (2000)	DP-FT-GENERATOR	GPT-2	47.5	32.0	32.1	32.0	31.9	32.0	32.1	32.0
Synthetic (3000)	DP-FT-GENERATOR	GPT-2	48.0	32.0	34.1	32.0	33.6	32.0	33.6	32.0
Synthetic (5000)	DP-FT-GENERATOR	GPT-2	48.3	35.8	32.7	32.0	30.5	32.0	35.6	31.1
Synthetic (2000)	AUG-PE	GPT-2	42.4	32.1 ↑	39.9 ↑	32.1 ↑	38.8 ↑	32.1 ↑	37.6 ↑	32.0
Synthetic (3000)	AUG-PE	GPT-2	43.2	32.0	39.1 ↑	32.0	38.6 ↑	32.1 ↑	39.5 ↑	32.1 ↑
Synthetic (5000)	AUG-PE	GPT-2	43.4	32.1	40.1 ↑	32.0	39.2 ↑	32.0	37.9 ↑	32.0 ↑
Synthetic (2000)	DP-FT-GENERATOR	GPT-2-Medium	49.7	36.5	40.3	32.0	33.5	31.9	35.6	31.9
Synthetic (3000)	DP-FT-GENERATOR	GPT-2-Medium	50.6	38.7	38.4	32.0	36.5	31.3	33.1	30.6
Synthetic (5000)	DP-FT-GENERATOR	GPT-2-Medium	50.3	41.2	39.8	31.4	37.4	31.7	34.6	31.0
Synthetic (2000)	AUG-PE	GPT-2-Medium	41.0	32.3	36.9	32.0	36.0 ↑	32.0 ↑	36.6 ↑	32.1 ↑
Synthetic (3000)	AUG-PE	GPT-2-Medium	42.1	32.1	38.3	32.1 ↑	38.9 ↑	32.1 ↑	37.5 ↑	32.1 ↑
Synthetic (5000)	AUG-PE	GPT-2-Medium	43.5	32.5	37.5	32.0 ↑	35.5	32.0 ↑	36.8 ↑	32.1 ↑
Synthetic (2000)	DP-FT-GENERATOR	GPT-2-Large	48.3	42.9	38.9	33.7	40.4	33.6	38.6	32.2
Synthetic (3000)	DP-FT-GENERATOR	GPT-2-Large	49.8	43.7	41.3	33.9	42.8	31.6	38.2	32.7
Synthetic (5000)	DP-FT-GENERATOR	GPT-2-Large	52.5	44.5	42.0	34.2	41.7	34.9	40.1	32.8
Synthetic (2000)	AUG-PE	GPT-2-Large	42.1	32.1	38.8	32.0	38.4	32.0	38.1	32.0
Synthetic (3000)	AUG-PE	GPT-2-Large	44.0	32.1	39.7	32.2	38.4	32.1 ↑	36.4	32.0
Synthetic (5000)	AUG-PE	GPT-2-Large	44.1	32.1	39.3	32.1	39.5	32.1	37.4	32.1
Synthetic (2000)	AUG-PE	GPT-3.5	45.4	43.5	43.5	44.6	42.8	44.5	41.9	43.1

C.11.1. COMPARISON TO WORD-LEVEL METRIC-DP FRAMEWORKS

Madib (Feyisetan et al., 2020) and TEM (Carvalho et al., 2023) employ metric differential privacy to privatize each word independently and achieve word-level ϵd-Metric DP, where d is the distance metric for neighboring words. Specifically, they perturb the embedding of each word and replace the current word with a new word whose embedding is closest to the noisy embedding. However, AUG-PE focuses on generating synthetic datasets with stronger guarantees provided by standard (ϵ, δ)-DP. Due to the fundamental differences in privacy definition: **(1) metric-DP v.s. DP; (2) word-level v.s. dataset-level privacy**, directly comparing our work with word-level metric-DP frameworks (Feyisetan et al., 2020; Carvalho et al., 2023) is not feasible.

To understand their privacy-utility tradeoff, we run Madib (Feyisetan et al., 2020) to generate samples under word level metric-DP with a high privacy budget $\epsilon = 10$. We followed their approach of perturbing 50-dimensional Euclidean GloVe embeddings with Laplace noise. We are unable to evaluate TEM (Carvalho et al., 2023) given that its code is not open-sourced.

Tb. 35 shows randomly sampled generated sentences from Madib and AUG-PE. Even with a high metric-DP budget ($\epsilon = 10$), Madib struggles to generate meaningful sentences on Yelp, OpenReview, and PubMed datasets. In contrast, AUG-PE, with a low DP budget ($\epsilon = 1$), can leverage GPT-3.5 as a data generator to produce fluent sentences across all three datasets.

C.11.2. COMPARISON TO SAMPLE-LEVEL LOCAL-DP FRAMEWORKS

Paraphraser (Mattern et al., 2022b) and DP Prompt (Utpala et al., 2023) focus on generating paraphrases for each private sample by varying the temperature during token sampling, which is regarded as a form of noise injection under the Local DP (LDP) framework. The sensitivity of each sample to the output can be constrained by clipping the logits of each generated token. While innovative, these methods' privacy budget scales linearly with the output's token length, presenting a challenge for generating longer sequences under a meaningful privacy budget.

Table 32: Next word prediction accuracy of downstream BERT$_{\text{Mini}}$ model under $\epsilon = \infty, 4, 2, 1$ on PubMed. (i) Compared to DP-FT-Generator, Aug-PE with a strong LLM GPT-3.5 can achieve higher accuracy under DP with the same synthetic data size. (ii) Compared to DP-FT-Downstream, Aug-PE can also obtain higher accuracy under $\epsilon = 2, 1$.

Data Type (Size)	Method	Data Generator	$\epsilon = \infty$ Accuracy	$\epsilon = 4$ Accuracy	$\epsilon = 2$ Accuracy	$\epsilon = 1$ Accuracy
Original (75316)	Fine-tune	-	43.5	30.7	28.9	26.7
Original (2000)	Fine-tune	-	33.5	2.2	1.8	1.4
Synthetic (2000)	DP-FT-Generator	GPT-2	30.2	27.8	27.6	27.2
Synthetic (3000)	DP-FT-Generator	GPT-2	31.1	28.7	28.4	28.1
Synthetic (5000)	DP-FT-Generator	GPT-2	32.4	29.7	29.4	29.2
Synthetic (2000)	Aug-PE	GPT-2	24.5	24.7	24.7	24.3
Synthetic (3000)	Aug-PE	GPT-2	25.7	25.6	25.4	25.0
Synthetic (5000)	Aug-PE	GPT-2	26.7	26.6	26.2	25.7
Synthetic (2000)	DP-FT-Generator	GPT-2-Medium	31.0	28.4	28.1	27.8
Synthetic (3000)	DP-FT-Generator	GPT-2-Medium	32.0	29.2	29.1	28.8
Synthetic (5000)	DP-FT-Generator	GPT-2-Medium	33.4	30.5	30.4	29.9
Synthetic (2000)	Aug-PE	GPT-2-Medium	25.5	25.4	25.1	24.9
Synthetic (3000)	Aug-PE	GPT-2-Medium	26.4	26.4	26.1	25.7
Synthetic (5000)	Aug-PE	GPT-2-Medium	28.0	27.6	26.9	26.1
Synthetic (2000)	DP-FT-Generator	GPT-2-Large	31.0	29.2	29.2	28.9
Synthetic (3000)	DP-FT-Generator	GPT-2-Large	32.2	30.3	30.1	29.8
Synthetic (5000)	DP-FT-Generator	GPT-2-Large	33.5	31.5	31.4	31.1
Synthetic (2000)	Aug-PE	GPT-2-Large	25.7	25.8	25.5	25.1
Synthetic (3000)	Aug-PE	GPT-2-Large	26.8	26.8	26.3	25.7
Synthetic (5000)	Aug-PE	GPT-2-Large	28.2	27.8	27.3	26.1
Synthetic (2000)	Aug-PE	GPT-3.5	30.4	30.3	30.2	30.1

It is worth noting that the mechanism for Local DP (taking a sample as input) and the mechanism for DP (taking a dataset as input) are not directly compatible. To establish a fair comparison between the Local DP in (Mattern et al., 2022b; Utpala et al., 2023) and DP employed by Aug-PE, we leveraged the conversion methodology in Feldman et al. (2022) to convert (ϵ_0)-LDP mechanism to (ϵ, δ)-DP mechanism for $\epsilon \ll \epsilon_0$, which requires shuffling the LDP outputs from each sample.

We use the code implementation provided by Feldman et al. (2022).[14] Due to the constraint that $\epsilon \ll \epsilon_0$,[15] the maximal Local DP ϵ_0 that can be used for a valid conversion on Yelp (with 1.9M private samples) is $\epsilon_0 = 8.785$, which corresponds to DP $\epsilon = 1.10$.

According to the Local DP guarantee in (Mattern et al., 2022b; Utpala et al., 2023), $\epsilon_0 = 2 * \texttt{n\_tokens} * (b_2 - b_1)/\texttt{temperature}$, where b_2, b_1 is the upper/lower bound for each token logit. We set $b_2 = 1$ and $b_1 = 0$ following (Mattern et al., 2022b). With $\texttt{temperature} = 2$, $\epsilon_0 = 8.785$ only allows generating $\texttt{n\_tokens} = 8$ tokens, which significantly hurts the utility of generated texts. To generate $\texttt{n\_tokens} = 64$ tokens for Yelp, one would need at least LDP $\epsilon_0 = 64$ under $\texttt{temperature} = 2$, and LDP $\epsilon_0 = 128$ under $\texttt{temperature} = 1$, which far exceeds practical limits for meaningful privacy guarantees. This contrasts with Aug-PE's capability to generate over $\texttt{n\_tokens} = 1000$ tokens while maintaining high quality under tight DP budgets (e.g., $\epsilon = 1$) in our experiments.

Furthermore, our attempt to directly evaluate Paraphraser (Mattern et al., 2022b) and DP Prompt (Utpala et al., 2023) was hindered by several practical challenges.

1. Paraphraser: The dataset used for finetuning in Paraphraser is not publicly available, and the implementation details necessary for replicating the exact privacy guarantees are absent.

2. DP Prompt does not specify the exact ϵ_0 used in the paper, focusing instead on empirical privacy attack accuracy as a proxy. The epsilon values are reported for all other baselines but not for Paraphraser and DP Prompt in Section 4.1

[14] https://github.com/apple/ml-shuffling-amplification

[15] https://github.com/apple/ml-shuffling-amplification/blob/993d285a546114bf8c70c33d053dca322a755707/computeamplification.py#L160

Table 33: Next word prediction accuracy of downstream $\text{BERT}_{\text{Small}}$ model under $\epsilon = \infty, 4, 2, 1$ on PubMed. (i) Compared to DP-FT-Generator, Aug-PE with a strong LLM GPT-3.5 can achieve higher accuracy under DP with the same synthetic data size. (ii) Compared to DP-FT-Downstream, Aug-PE can also obtain higher accuracy under small privacy budget.

Data Type (Size)	Method	Data Generator	$\epsilon = \infty$	$\epsilon = 4$	$\epsilon = 2$	$\epsilon = 1$
			Accuracy	Accuracy	Accuracy	Accuracy
Original (75316)	Fine-tune	-	47.6	34.1	32.5	30.4
Original (2000)	Fine-tune	-	34.6	1.1	0.8	0.6
Synthetic (2000)	DP-FT-Generator	GPT-2	32.4	29.7	29.4	29.2
Synthetic (3000)	DP-FT-Generator	GPT-2	33.1	30.5	30.3	30.0
Synthetic (5000)	DP-FT-Generator	GPT-2	34.3	31.4	31.2	30.9
Synthetic (2000)	Aug-PE	GPT-2	26.7	27.0	26.9	26.5
Synthetic (3000)	Aug-PE	GPT-2	27.7	27.6	27.6	27.3
Synthetic (5000)	Aug-PE	GPT-2	28.5	28.5	28.3	27.9
Synthetic (2000)	DP-FT-Generator	GPT-2-Medium	33.1	30.2	30.0	29.8
Synthetic (3000)	DP-FT-Generator	GPT-2-Medium	33.8	31.3	30.9	30.6
Synthetic (5000)	DP-FT-Generator	GPT-2-Medium	35.2	32.1	32.1	31.7
Synthetic (2000)	Aug-PE	GPT-2-Medium	27.7	27.6	27.4	27.0
Synthetic (3000)	Aug-PE	GPT-2-Medium	28.5	28.5	28.3	27.7
Synthetic (5000)	Aug-PE	GPT-2-Medium	29.8	29.6	28.9	28.4
Synthetic (2000)	DP-FT-Generator	GPT-2-Large	33.1	31.2	31.1	31.1
Synthetic (3000)	DP-FT-Generator	GPT-2-Large	34.2	32.4	32.2	32.0
Synthetic (5000)	DP-FT-Generator	GPT-2-Large	35.4	33.5	33.2	33.0
Synthetic (2000)	Aug-PE	GPT-2-Large	27.9	27.9	27.7	27.2
Synthetic (3000)	Aug-PE	GPT-2-Large	28.9	28.8	28.5	27.7
Synthetic (5000)	Aug-PE	GPT-2-Large	30.2	29.8	29.3	28.3
Synthetic (2000)	Aug-PE	GPT-3.5	32.7	32.5	32.5	32.4

of Utpala et al. (2023). Additionally, as mentioned in Section 4.3 of Utpala et al. (2023), models like ChatGPT do not expose logits, so the authors do not perform a logit clipping operation in many of their experiments. This further disables the computation of an exact ϵ_0 and renders a direct quantitative comparison between DP Prompt and Aug-PE infeasible.

C.12. Aug-PE Convergence under One Private Sample

In this section, we only use *one* private example in Alg. 1 to generate *one* synthetic sample. We qualitatively examine if the synthetic sample from Aug-PE increasingly resembles this specific private sample over the PE iterations. This offers a clearer illustration of Aug-PE's convergence behavior. Specifically, at each iteration, we generate K variations for the current synthetic sample, use the private sample to identify and vote for its nearest synthetic sample based on their embeddings, and select the nearest synthetic sample for the next iteration. Tb. 36 and Tb. 37 show the generations results from GPT-3.5 under one Yelp private sample and one OpenReview private sample, respectively.

As shown Tb. 36, after the voting, the selected synthetic sample relates to the term "taco", a word present in the private example. By the second iteration, the synthetic sample includes the term "Mexican food", which aligns with the central theme of the private example. By the fifth iteration, the phrase "authentic Mexican food" surfaces in the synthetic sample, resonating with phrases like "real deal Mexican food" and "great authentic food" from the private example. This demonstrates that the synthetic sample increasingly aligns with the private sample as the iterations progress.

In the OpenReview example presented in Tb. 36, we note that the initial synthetic sample at iteration 0 pertains to the privacy aspects of machine learning, whereas the private sample focuses on adversarial detection and robustness. As the iterations progress, by iteration 4, the topic of synthetic sample shifts to "inference attack in machine learning", which aligns with the robustness theme of the private sample. By the fifth iteration, terms like "Adversarial Attacks in Machine Learning" and "Robustness-Enhancing" emerge in the synthetic sample, similar to the topic of "adversarial detection" from the private sample. It shows that the synthetic sample shifts the topic from privacy to robustness over PE iterations, progressively aligning more closely with the private sample.

The above two examples demonstrate that Aug-PE can converge, by producing diverse variations and effectively selecting

Table 34: A qualitative comparison between AUG-PE and text-to-text privatization approaches.

Name	Method	Source of Randomness	Sensitivity Control	Privacy Guarantee
Madib (Feyisetan et al., 2020) TEM (Carvalho et al., 2023)	Word embedding perturbation	Word Embedding Noise	N/A (metric distance is included in privacy definition)	Word-level metric-ϵd DP where d is the distance metric
Paraphraser (Mattern et al., 2022b) DP Prompt (Utpala et al., 2023)	Paraphrasing with temperature	Temperature in the next token sampling stage	Clipped logits of each token	(Sample-level) ϵ Local DP
AUG-PE (Ours)	Private evolution	Histogram noise	Each private sample only contributes one vote in the histogram	(Dataset-level) standard (ϵ, δ)-DP

Table 35: Randomly sampled synthetic data from Madib (Feyisetan et al., 2020) (word level metric-DP $\epsilon = 10$) and AUG-PE (DP $\epsilon = 1$). AUG-PE with data generator GPT-3.5 yields higher quality texts.

Method	Privacy Guarantee	Yelp	OpenReview	PubMed
Madib	Word level metric-DP $\epsilon = 10$	i was born including raised he hardwick create during combination school , tony jones was in 'go to ' place for the greatest pizza ever . n't do n't live completed whitley rich and crazy visit a preparing times a year with weeks night cut 6 civil us took to tony 's work dinner bogota of normal end my 'local ' strangers , tried would suggest instance visitors ? seen 'd was yet to gone . . . (**omitted**)	. paper proposed bringing reinforcement buddhism based approach money automatically predictions graph augmentations for a graph neural network (gnn) classification problem there few authors creates taken label invariance (data augmentations that do protect risks labels) is part rich also felonies problem dealt gnn partner with. . . (**omitted**)	mandibular overdentures many a selection treatment option for placed edentulous diabetes only long-term predictable outcomes , using suspension loading facilitated cone currently , could early well repatriation loading protocols same mandibular implant overdentures number prevalent in in literature details a systematic review ,... (**omitted**)
AUG-PE (GPT-3.5)	DP $\epsilon = 1$	The fried chicken and the collard greens were some of the best Southern fare we've ever had, not to mention the amazing gumbo. We highly recommend this restaurant if you're in the area and can't wait to try some of their other flavorful dishes. Everything filled us up and left us satisfied.	This paper presents an innovative method of deep representational learning for facial expression recognition. The method is evaluated on CIFAR-10 and ImageNet datasets and it is comprehensive encompassing all facets of saliency modeling to proposed deep representational features for representing multiple saliencies. The paper is in a well-structured and the methods are clear. . . (**omitted**)	In this retrospective study, we aimed to investigate the prevalence of stroke and identify the factors associated with its occurrence. Data were extracted from medical records, along with symptoms, electrocardiograms (ECGs), and syncope in a cohort of patients with a mean age of 71 years. Of the total 345 patients, 28% had cardiac abnormalities as revealed by ECGs, significantly higher than those without [p<0.001].. (**omitted**)

ones that closely align with the private example.

Table 36: Generation results from Aug-PE under *one* Yelp private sample. The synthetic sample generated from GPT-3.5 aligns more closely with the training sample (Yelp review about Mexican food) over iterations. The text length of the synthetic sample is controlled through our adaptive length mechanism. Keywords with semantic similarity to the private sample are highlighted.

Private	Real deal Mexican food in Chadds Ford, PA. Excellent burritos, tacos, and tamales. I plan to be a regular. Meals served traditionally with sliced radishes, pumpkin seeds and lime. Freshly prepared fillings, onion, cilantro, and red and green sauces. Brightly colored building across from Bavarian Motors. Nothing fancy, but great authentic food.
iter 0	I recently visited this restaurant having heard that they had one of the best taco stands in town. I was not disappointed! The tacos were incredibly delicious and tasted like they were made with fresh ingredients. The portions were generous, and the prices were extremely reasonable. The service was excellent too, with the staff being attentive and accommodating. The ambience of the restaurant was also great, making it a perfect spot for a casual night out. Overall, I had a fantastic experience and would definitely recommend this place to all those who love tacos. A definite 5-star rating!
iter 1	Despite having heard that they had one of the best taco stands in town, I was not disappointed. The tacos were incredibly delicious and tasted like they were made with fresh ingredients. The portions were generous, and the prices were reasonable. The staff was attentive and accommodating. The ambience was great, making it a perfect spot for a casual night out. Overall, I had a fantastic experience and would highly recommend this place to all those craving Mexican cuisine. A solid 5-star rating from me!
iter 2	Despite having heard that they had some of the best Mexican food in town, the crowded parking lot stands in contrast. However, that was not the case. The food tasted authentic and fresh, with the guacamole and salsa being particularly noteworthy. The portions were generous, and the prices were reasonable. The staff was attentive and accommodating. The ambience was great, making it a perfect spot for a date night or a family dinner. Overall, I had a fantastic experience and would highly recommend this
iter 3	Many had heard that they had amazing Mexican food in town, but the crowded parking lot stands in contrast. However, once you step inside, the experience is incredible. The food is authentic, and the guacamole and salsa being particularly noteworthy. Prices were as well quite reasonable. The staff was attentive and accommodating. The ambiance was great, making it perfect for a date night or family dinner. I had a fantastic experience and highly recommend the restaurant to anyone seeking quality Mexican cuisine.
iter 4	Many locals had heard that they had the best Mexican food in town and the hype stands in its truth. However, upon stepping into the experience, it was incredible. The food was fresh, flavorful, and authentic with the guacamole and salsa being particularly noteworthy. The portions were well-sized and satisfying. The staff was attentive and accommodating. The ambiance was cozy and intimate, making it perfect for a romantic night out or casual dinner with friends. I had a great time and highly
iter 5	Many locals had raved about the best authentic Mexican food in town and they were not exaggerating. Walking in to the restaurant was refreshing. The food was fresh, the guacamole and salsa were noteworthy. The margaritas were strong and satisfying. The staff were attentive and accommodating. The ambiance was cozy and intimate, making it perfect for a romantic dinner with loved ones. The prices were great and the portions were generous. The fajitas were sizzling and the tacos were packed with flavor. Overall, this
iter 6	Many locals have raved about the best authentic Mexican food in town and they were not exaggerating. Walking in the restaurant was refreshing. The decor was colorful, lively and festive. The margarita was perfectly balanced and satisfying. The waiter was attentive and accommodating. The ambiance was cozy and intimate, making it perfect for a romantic dinner. The chips and salsa were packed with flavor. Overall, the food was great and the portions were generous. The fajitas were sizzling and the chicken was tender. The guacam
iter 7	Many locals consistently rave about the amazing authentic Mexican cuisine and we were not disappointed. The flavors were refreshing, the decor colorful and lively, and the margaritas satisfying. The waiter was attentive and accommodating. The outdoor seating was perfect for a romantic dinner and the chips and salsa were packed with flavor. Overall, the food was top-notch and the portions were generous. The chicken enchiladas and guacamole were particularly noteworthy. It was a great experience and we highly recommend this restaurant to anyone looking for a delicious meal and a
iter 8	Many foodies rave about the amazing authentic Mexican cuisine and they were not wrong. The flavors were richly robust, colorful and enticing, and the margaritas were top-notch. The enchiladas and guacamole were particularly outstanding. The service was attentive and accommodating and the outdoor seating was perfect for a leisurely dinner and people-watching. From the chips and salsa to the flavorful entrees, the food was superb with generous portions. The menu offers a large selection of delectable options which makes it easy to recommend to anyone looking for a satisfying meal
iter 9	Many foodies have raved about the amazing authentic Mexican cuisine. The flavors were bold and delicious. The tacos were rich and flavorful, with enticing toppings and freshly made guacamole. The service was attentive and accommodating, and the outdoor seating was perfect for a warm summer evening. From the crispy chips and flavorful salsa to the perfectly spiced dishes, the food was generous and satisfying. The menu offers a large selection of vegetarian and meat options which makes it easy to please anyone's taste buds. Overall, the experience was excellent and highly recommended for people looking for an authentic Mexican dining experience.

Table 37: Generation results from Aug-PE under *one* OpenReview private sample. The topic (highlighted) of the generated sample changes from “privacy in machine learning” (iter 0) to “adversarial robustness” (iter 6), which is more closer to “adversarial detection” topic in the private sample. This shows that the synthetic sample generated from GPT-3.5 aligns more closely with the training sample over iterations. The text length of the synthetic sample is controlled through our adaptive length mechanism.

Private	The authors propose a new statistic, called Expected Perturbation Score (EPS), for adversarial detection. Based on EPS, the authors develop a Maximum Mean Difference (MMD) metric to measure the difference between test samples and natural samples, and further propose an EPS-based adversarial detection method (EPS-AD). Sufficient theoretical analysis and extensive experiments demonstrate the correctness and effectiveness of the proposed method. Strength: 1. This paper is easy to read. 2. The authors propose a new statistical method, called Expected Perturbation Score (EPS), which is able to obtain enough information to identify adversarial examples with only one example after various perturbations. 3. Sufficient theoretical analysis is performed to demonstrate that EPS is able to simulate the difference between the two distributions under mild conditions. Furthermore, extensive experimental results demonstrate the superiority of the proposed EPS-AD. 4. The proposed EPS will be an effective statistic in many applications, such as out-of-distribution detection and anomaly detection. Weaknesses: This paper is easy to read and interesting but still has some minor issues, please refer to weaknesses.
iter 0	Review Example: Overall, this paper provides a comprehensive discussion on privacy issues related to machine learning applications. The authors discuss the concept of differential privacy as a potential solution to protect individual privacy while still maintaining the accuracy of the model. They also provide examples of implementation and evaluation of privacy-preserving techniques using real-world datasets. Strengths: - The paper provides a thorough explanation of differential privacy and how it can be utilized in various machine learning techniques. - The use of real-world datasets to demonstrate implementation and evaluation of privacy techniques adds credibility to the work. - The paper highlights the importance of privacy in machine learning and provides readers with potential solutions to protect individual privacy. Weaknesses: - More evaluation of the privacy techniques could be provided to better understand their effectiveness. - The paper focuses primarily on differential privacy and does not explore other privacy-preserving techniques. - The practical implementation of the techniques may be complex for those without a strong understanding of machine learning and data analysis. In summary, this paper is a strong contribution to the area of social aspects of machine learning, specifically in the field of privacy. It analyzes the concept of differential privacy, its implementation, and evaluation to protect individual privacy in machine learning models. Despite some limitations, the paper contributes significantly to the field and is recommended for acceptance.
iter 1	In their paper ’Privacy in Machine Learning: An Overview’, the authors provide a comprehensive discussion on privacy issues in machine learning. The authors discuss the concept of differential privacy as a potential solution for individual privacy while maintaining the utility of the data. They also provide examples and analysis of privacy-preserving techniques on real-world datasets such as the Stanford and MNIST datasets. The thorough explanation of differential privacy and how it can be utilized in various machine learning techniques adds credibility to the paper. The paper highlights the importance of privacy in machine learning and provides a potential solution to individual privacy. Weaknesses: More evaluation of the privacy techniques could be provided to better understand their effectiveness. The paper primarily focuses on differential privacy and does not explore other privacy-preserving techniques. The technical implementation of the techniques may be difficult for those without a strong understanding of machine learning and statistical analysis. Overall, this paper contributes significantly to the area of social aspects of machine learning, and it is recommended for acceptance as a good paper.
iter 2	In this paper, titled ’Policy Machine Learning: An Overview of the Discussion on Privacy Issues in Machine Learning’, the authors discuss privacy as a potential concern for individual privacy while maintaining the integrity of the data. They also provide examples of privacy-preserving techniques on real-world data from Stanford and MNIST. The thorough exploration of differential privacy can be utilized effectively which adds credibility to the paper. The paper emphasizes the importance of privacy in machine learning and provides a valuable contribution to the field. Weaknesses include the evaluation of techniques to be used to assess their effectiveness. The paper focuses on privacy issues and does not explore fairness-preserving methods. With its contribution to the social aspects of machine learning and statistical analysis, the paper is recommended with a rating of 8 as a good paper.
iter 3	In their research paper, Inference Attack Policy Machine Learning: An Interpretable and Almost True Framework for Predictive Analytics, the authors highlight potential concerns for individual privacy while discussing the importance of privacy in machine learning. They also provide examples of how sensitive data from ImageNet and MNIST datasets can be utilized effectively while ensuring thorough differential privacy which adds credibility to the paper. The research emphasizes the importance of interpretability in machine learning, making a valuable contribution to the field of social aspects of machine learning. We recommend including case studies of how interpretability can be used to assess their effectiveness. The paper also outlines how it does not explore fairness and ethics methods. With this contribution to the field of machine learning and statistical modeling, the authors provide a valuable framework for policy inference attack in machine learning.
iter 4	In their research paper, ’Inference Attacks in Machine Learning: An Interpretability and Almost Interpretability Framework and its Application to Privacy and Analytics’, the authors highlight the need for protecting sensitive data in machine learning. They provide examples of sensitive data from ImageNet and NIST datasets, emphasizing the importance of being thorough in privacy protection to ensure credibility to their research. The paper stresses the importance of interpretability in machine learning. By making a valuable contribution to this field, it provides case studies of how interpretability can be used to assess the effectiveness of machine learning models. The paper outlines various approaches to exploring fairness, transparency, and ethics in machine learning. The results of the study contribute to the need for a comprehensive policy to prevent inference attacks in machine learning.
iter 5	In our research paper, entitled ’Adversarial Attacks in Machine Learning: An Interpretability-Almost-Explainability Framework and its Application to Private Data Analysis’, the authors emphasize the need for protecting sensitive data in machine learning. They provide examples by using data from Inet and MNIST dataset, and address the importance of privacy to ensure the credibility of the results. The paper is well-written and well-structured, making a valuable contribution to the field. Additionally, it highlights the importance of interpretability to enhance the effectiveness of machine learning models. The paper also focuses on fairness, transparency, and ethics in machine learning and the study presents a comprehensive analysis in adversarial attacks. We highly recommend accepting this good paper.
iter 6	In our research paper titled ’Adversarial Attacks in Machine Learning: An Interpretable and Robustness-Enhancing Framework and Empirical Data Analysis’, the authors emphasize the significance of interpretability in machine learning. They provide a comprehensive approach using Integrated Gradients and M-Taylor expansions, to address the challenges and ensure the robustness of results. The paper is well-written, making valuable contributions to the field, and emphasizes the importance of interpretability to enhance the effectiveness of machine learning. Moreover, the study presents a comprehensive approach in defending against adversarial attacks. Therefore, I recommend accepting this good paper.