DPGA-TextSyn: Differentially Private Genetic Algorithm for Synthetic Text Generation

Zhonghao Sun[1], Zhiliang Tian[1*], Yiping Song[2], Yuyi Si[1], Juhua Zhang[1], Minlie Huang[3], Kai Lu[1], Zeyu Xiong[1], Xinwang Liu[1], Dongsheng Li[1*]

[1]College of Computer Science and Technology, National University of Defense Technology

[2]College of Science, National University of Defense Technology

[3] Department of Computer Science, Tsinghua University

{ sunzhonghao, tianzhiliang, songyiping, siyuyi, zhangjuhua23, kailu, xiongzeyu08, xinwangliu, dsli }@nudt.edu.cn

aihuang@tsinghua.edu.cn

Abstract

Using large language models (LLMs) has a potential risk of privacy leakage since the data with sensitive information may be used for fine-tuning the LLMs. Differential privacy (DP) provides theoretical guarantees of privacy protection, but its practical application in LLMs still has the problem of privacy-utility trade-off. Researchers synthesized data with strong generation capabilities of closed-source LLMs (i.e., GPT-4) under DP to alleviate this problem, but this method is not so flexible in fitting the given privacy distributions without fine-tuning. Besides, such methods can hardly balance the diversity of synthetic data and its relevance to target privacy data without accessing so much private data. To this end, this paper proposes DPGA-TextSyn, combining general LLMs with genetic algorithm (GA) to produce relevant and diverse synthetic text under DP constraints. First, we integrate the privacy "gene" (i.e. metadata) to generate better initial samples. Then, to achieve *survival of the fittest* and avoid *homogeneous*, we use privacy nearest neighbor voting and similarity suppression to select elite samples. In addition, we expand elite samples via genetic strategies such as mutation, crossover, and generation to expand the search scope of GA. Experiments show that this method significantly improves the performance of the model in downstream tasks while ensuring privacy.

1 Introduction

Using large language models (LLMs) in sensitive textual domains (Schmiedmayer et al., 2024) poses significant privacy risks, where attackers may hack the fine-tuned LLMs to extract privacy information in the LLMs' fine-tuning data (Carlini et al., 2021). To prevent privacy data from being extracted, researchers apply differential privacy (DP) (Dwork, 2006) to LLMs, providing a theoretical guarantee for protection by adding calibrated noises. However, its core challenge is the privacy-utility trade-off: While DP offers strong privacy guarantees, model utility often suffers from overly conservative noise [1]. So, enhancing LLMs' utility under DP is a key topic.

Researchers explored different ways to implement DP, mainly in two categories. (1) Model-free methods add calibrated noise to the input/output to anonymize text (Feyisetan et al., 2020; Wu et al., 2023), which does not require model fine-tuning and can adapt to closed-source LLMs like GPT-4. However, since the noise must be added to each sample, the noise scale increases with the length and number of samples; and the increasing of noise reduces the model utility. (2) Model-based methods feed noises into the model training (Abadi et al., 2016; Yu et al., 2021). The advantage is that once the model training is accomplished, privacy loss is fixed (stopping adding additional noise). Therefore, the strength of privacy protection and utility does not decrease as the users feed new samples to the model for inference. However, these methods require sufficient domain-specific supervised training samples, which are often hard to obtain (Breuer et al., 2020) in data-sensitive fields such as medicine. Additionally, since DP fine-tuning adds noise to gradients, the privacy cost increases with more training data. Ensuring reasonable privacy guarantees requires adding more noise, thus degrading model utilities.

To further enhance model-based methods, researchers used LLMs' generative abilities to synthesize training samples under DP. These methods construct the DP-based generators by supervised fine-tuning or prompt-tuning LLMs to produce samples for downstream tasks. According to the DP post-processing property (Dwork et al., 2014),

*Corresponding Author

[1]DP always requests noises based on the worst case.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16159–16179

once synthetic data are generated, no additional noise needs to be required to account for privacy loss. So, downstream tasks can use almost infinite samples from the generator without slashing privacy protection. There are two main categories of work using DP to synthesize text. (1) Someone fine-tune LLMs under DP to synthetic data (Yue et al., 2022; Yu et al., 2023). This method inherits DP fine-tuning limitations, including requiring many supervised privacy samples. Moreover, they require trainable LLMs, so it is unsuitable for closed-source LLMs like GPT-4 [2]. (2) Methods in the second category use LLMs APIs to synthesize text similar to privacy data (Song et al., 2024; Xie et al., 2024). This method does not require large-scale domain-specific data and is adaptable to closed-source LLMs. Compared with fine-tuned LLMs, closed-source LLMs are typically stronger in synthesizing text but are not so flexible in fitting the given privacy distributions due to the lack of direct fine-tuning. It is hard to balance the diversity of synthetic data and its relevance to target privacy data. Low relevance leads to a gap between synthetic data and data required by downstream tasks (i.i.d. with the privacy set); low diversity makes the model hard to cover the whole privacy distribution and thus the downstream models may overfit partial privacy distribution.

To synthesize high-quality samples, we should ensure the synthesized samples are both diverse and relevant to the privacy data. Genetic algorithm can effectively evolve populations, which is suitable for our scenarios: our model may inherit the "gene" information from the privacy data and evolve to obtain new data maintaining the "gene" while keeping the diversity.

In this paper, we propose **D**ifferential **P**rivacy **G**enetic **A**lgorithm for **Text Syn**thesis method (**DPGA-TextSyn**)[3] to protect the privacy of sensitive data, which imports the idea of genetic evolution to LLMs to generate high-utility synthetic text under DP constraints. The core innovation is to synthesize samples according to "gene" of the privacy data and expand more samples via genetic evolution mechanism. Specifically, to make synthetic samples consistent with the privacy data, we extract the "gene" (i.e. metadata) of privacy data under DP protection to obtain the initial synthetic samples (§3.2). To achieve *survival of the fittest* and avoid *homogeneous* per generation, we select distinct samples to construct an elite set for each generation. Here, we construct a DP histogram based on nearest neighbor voting to select elite samples and employ similarity suppression to avoid samples being homogeneous (more and more similar) as iterations (§3.3). To evolve to obtain more samples, we design three genetic strategies including mutation, crossover, and generation, to expand the elite set, which ensures the quality and diversity in the next generation. (§3.4). Experiments show that DPGA-TextSyn excels baselines.

Our contributions are: (1) We propose DPGA-TextSyn, a method that implements genetic algorithms to synthesize data under DP, it uses LLMs APIs to iteratively generate synthetic text data that is close to privacy data, addressing privacy-utility trade-offs in data-scarce scenarios. (2) We propose to extract privacy metadata "gene" to generate better initial synthetic data with a low privacy cost, which instantiates SVT to obtain statistics and conduct privacy voting to obtain meta descriptions. (3) Experiments show our model excels strong baselines in downstream tasks.

2 Preliminaries and Related work

2.1 Preliminaries

Definition 2.1 (Differential Privacy(Dwork et al., 2014)). A randomized algorithm $\mathcal{M}$ is (ε, δ)-differential privacy (DP) if for any pair of neighboring dataset D and D', and any $S \subseteq \text{Range}(\mathcal{M})$,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

ε measures the privacy loss. δ allows for a small probability that the privacy guarantee may fail.

2.2 Related work

DP for LLMs. As LLMs grow more prevalent, privacy attacks have become increasingly sophisticated, highlighting increasing privacy risks in LLMs (Li et al., 2023; Kojima et al., 2022). Differential Privacy (DP) (Dwork, 2011) is used across machine learning stages, including input, training, and inference. At the input stage, DP ensures privacy by perturbing data features. Feyisetan et al. replaced the sensitive text with semantically similar words in the embedding space (Feyisetan et al., 2019). At the training stage, gradient-based DP

[2]Some closed-source LLMs offer fine-tuning API, but dp fine-tuning needs special implementation in model optimization. Currently, no model provides custom API.

[3]The code for our method is available at: https://github.com/szzhh/DPGA-TextSyn

like DP-SGD provides privacy by clipping gradients and adding Gaussian noise (Abadi et al., 2016). DP applications in inference focus on privacy protection in output generation (Lee and Kifer, 2018). Wu et al. proposed the Report-Noisy-Max mechanism, adding noise to inference results for privacy (Wu et al., 2023).

DP synthetic text is generated by training models to produce new text similar to real data while protecting personal information (Mattern et al., 2022). Yue and Putta et al. explored using DP-SGD to fine-tune pre-trained language models like GPT-2 for generating synthetic text datasets (Yue et al., 2022; Putta et al., 2022). Mattern and Kurakin et al. demonstrated that training downstream models on DP synthetic text can achieve performance comparable to DP training directly on real data (Mattern et al., 2022; Kurakin et al., 2023). However, achieving a good trade-off between fidelity and privacy requires large batch sizes and long training iterations (Anil et al., 2021). The rise of closed-source LLMs like GPT-4 has made DP fine-tuning infeasible, driving research on API-based generation methods (Touvron et al., 2023). A new approach uses API access to pre-trained models and applies DP to select samples similar to privacy data (Zhang et al., 2024), generating variations of these samples (Lin et al., 2023). The AUG-PE algorithm improves generation and selection via APIs to produce high-quality DP synthetic text (Xie et al., 2024).

Some additional related work is added in App. A.

3 Method

3.1 Overview

As Fig. 1, our model consists of three modules: (1) **Initial Sample Synthesizing** (§3.2) generates the initial version of synthetic data according to the privacy metadata (i.e. statistics and keywords) under a low privacy cost, where we use sparse vector technique and keywords selection to obtain the privacy metadata. (2) **Distinct Elite Sample Selection** (§3.3) employs private nearest neighbor voting and similarity suppression to obtain elite set by selecting synthetic data, ensuring that the diversity of synthetic data is more like that of privacy data. (3) **Elite Set Expansion** (§3.4) constructs three different genetic operations to expand the elite set (i.e. mutate, cross, and generate), which balance the quantity and quality of synthetic data.

To synthesize samples, we first generate initial synthetic samples (§3.2), vote to select the distinct elite set over initial samples (§3.3), and then expand the elite set (§3.4). At the same time, following the elitist strategy[4] of GA, we retain the elite set to the next generation. We repeat the operations in §3.3 and §3.4 for T times and obtain the final samples for downstream tasks.

3.2 Initial Sample Synthesizing via Privacy Metadata

We synthesize the initial samples via privacy metadata (statistics and key description information of privacy data), which ensures they better align with the privacy samples. The privacy cost of accessing full privacy data is high, so we obtain the privacy metadata at a very low privacy loss by adding Laplace noise[5] to the histogram. The histogram meets the application conditions of parallel combination (see in App. B), thus we can limit the DP noise within a privacy budget of ε, even if the counting result of each bin is added with DP noise with a privacy budget of ε (Dwork et al., 2014). The metadata consists of (1) statistical information (i.e. label distribution and length distribution) of privacy data for distribution close to privacy data. (2) key description of privacy data for semantics close to privacy data. We integrate the information mentioned above into the prompt to guide the LLMs to generate better initial synthetic samples.

3.2.1 Obtaining statistics as metadata

We obtain label and length distribution histograms through statistics, which are easily accessible, valuable, and robust to noise.

- **Label distribution.** According to the label distribution $Y = \{y_1, y_2, \ldots, y_k\}$ of privacy data , we construct a label histogram $H = \{h_1, h_2, \ldots, h_k\}$, where h_i represents the count of samples for label y_i . To satisfy DP, we add Laplace noise η_i to each count, forming a noisy histogram $\tilde{H} = (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_k)$, where $\tilde{h}_i = h_i + \eta_i$. Then, we use $\tilde{H}$ to guide the label distribution of synthetic data, which makes the synthesized data close to privacy data distribution with low privacy costs.

- **Text length distribution.** Since the length distributions of different privacy datasets are different and

[4]Elitist Strategy in GA preserves the optimal solution from each generation by copying it unchanged to the next, preventing loss due to crossover and mutation.

[5]Laplace noise is common in DP and is more suitable for SVT, so we add it in all operations that require noise in §3.2.

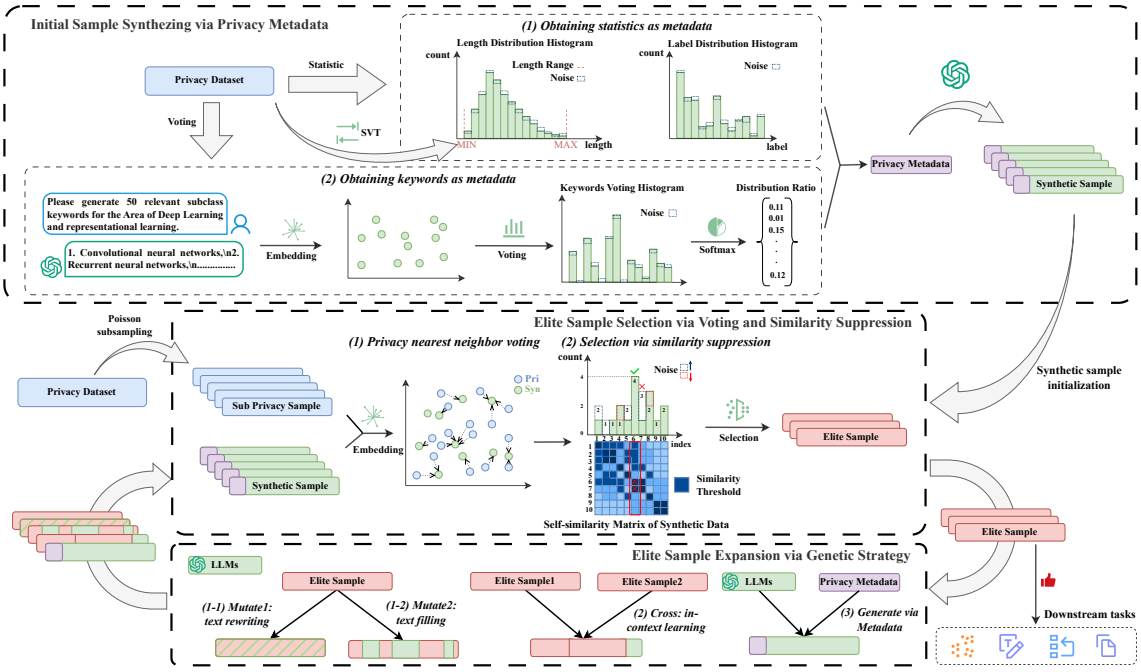

Figure 1: DPGA-TextSyn framework. Three black dotted boxes represent the specific implementation of the three algorithm sub-modules(§3.2, §3.3, §3.4).

are crucial in synthesis, we use the sparse vector technique (SVT) in DP theory (see in App. B.2) to get an approximate text length (i.e. number of tokens) range (i.e. minimum and maximum) of the privacy dataset. Then, we count the text lengths of all privacy samples in this range to construct a length distribution histogram and add calibrated noise to it. The specific steps for implementing SVT to get length distribution are as follows: (1) **Set the initial maximum length.** We preset the initial maximum length to 0 because it must be lower than the true privacy sample length maximum. (2) **Construct the query stream.** We gradually increase the preset maximum length with a step size of 1, and query the privacy dataset at each step under DP (add Laplace noise) to obtain the number of samples whose text lengths are greater than the current maximum length. (3) **Get the approximate maximum length.** We set the threshold of SVT to 0 (it ensures that no sample length exceeds the current maximum length), and add Laplace noise as the noise threshold. For each query result in step (2), if it is less than or equal to the noise threshold, we regard the maximum length corresponding to this step as the approximate maximum length because it is greater than the length of all samples. (4) **Generate the length distribution histogram.** Similarly, we get the minimum length. Within the range of the approximate minimum and maximum lengths, we count the length of the privacy data to generate the length distribution histogram and add Laplace noise to this histogram to satisfy DP. SVT costs a fixed privacy budget for all queries, which reduces the overall privacy cost.

3.2.2 Obtaining keywords as metadata

To get descriptions as privacy metadata, we employ LLMs to generate some keywords to describe the privacy data distribution and sample privacy data to act as DP-based discriminators to vote for the keywords. It consists of three steps: (1) We first **generate some keywords** from task-related public information. As we mainly focus on the classification task in this paper, we regard the class name as the public information (Yue et al., 2022) and prompt LLMs to generate some possible subclass names (keywords) as more specific information [6] (see prompts in App. I). (2) We then **generate a DP-noised histogram via voting** with the privacy data. Specifically, we use the nearest neighbor to vote for keywords like §3.3.1, count the votes to generate the histogram, and add Laplace noise to this histogram to satisfy DP. (3) Finally, we **discriminate keywords via noised histogram**. We apply the softmax function to this noised histogram to get the percentage of votes for each keyword,

[6] Our experiments also involve generation tasks, where no existing class name are available, we generate the subclass names directly by prompting LLMs.

based on which we discriminate to pick out good keywords for the next initial sample generation.

3.3 Distinct Elite Sample Selection via Privacy Voting and Similarity Suppression

To ensure that the evolution direction of the synthetic data distribution is gradually moving towards the privacy data, we propose a distinct elite sample selection mechanism via genetic algorithm (GA) to discard synthetic samples that are far from privacy data and select high-quality synthetic samples to "reproduce" the next generation. Specifically, (1) we construct a voting histogram via privacy nearest neighbors voting and iteratively select samples according to the number of votes from high to low, which ensures that the selected samples are relevant to the privacy data. (2) We use similarity suppression when selecting to avoid selecting samples that are very similar to the selected samples.

3.3.1 Privacy nearest neighbor voting

Let $\mathcal{D}_{pri}$ be the privacy dataset. We subsample $\mathcal{D}_{pri}$ and use it to vote for the synthetic data via the nearest neighbor method. Specifically, we first perform Poisson subsampling on $\mathcal{D}_{pri}$ to obtain a privacy subset $\mathcal{D}'_{pri}$. Subsequently, we use the text-embedding-ada-002[7] embedding model Φ to get the embeddings for $\mathcal{D}'_{pri}$ and the synthetic data $\mathcal{D}_{syn}$. Let $\mathbf{e}_i^{pri}$ be the embedding of the i-th sample in $\mathcal{D}'_{pri}$ and $\mathbf{e}_j^{syn}$ be the embedding of the j-th sample in $\mathcal{D}_{syn}$. Then, each private sample in $\mathcal{D}'_{pri}$ votes for the closest synthetic sample by calculating the cosine similarity $\text{Cosine}(\mathbf{e}_i^{pri}, \mathbf{e}_j^{syn})$ of their embeddings. Let v_j be the number of votes that the j-th synthetic sample receives. Finally, we count the votes of all synthetic samples, i.e., for each j, we have $v_j = \sum_i \mathbb{I}(\text{argmax}_k \text{Cosine}(\mathbf{e}_i^{pri}, \mathbf{e}_k^{syn}) = j)$ (where $\mathbb{I}$ is the indicator function), and add the calibrated Gaussian noise [8] $\mathcal{N}(0, \sigma^2)$ to voted results to meet DP. That is, the final count for the j-th synthetic sample is $v_j^{final} = v_j + \mathcal{N}(0, \sigma^2)$.

Note that the privacy voting here can transform the traditional DP fine-tuning generation paradigm into a DP selection discrimination paradigm, which greatly improves the problem of a tight privacy budget under the premise of ensuring utility because DP fine-tuning LLMs for generation requires adding noise to the gradient at the sample level, while DP selecting public samples from LLMs only requires adding noise to the histogram, which is simpler and requires less privacy budget.

3.3.2 Selection via similarity suppression

We construct a self-similarity matrix for similarity suppression, which is combined with the voting result above to select elite samples. The specific steps are as follows: (1) **Self-similarity matrix construction.** Suppose we have n synthetic samples, and their embedding vectors are denoted as $\mathbf{x}_i$, where $i = 1, 2, \cdots, n$. The elements S_{ij} of the self-similarity matrix S are calculated by the cosine similarity, here $S_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$. Similarly, we construct the self-similarity matrix P of privacy data[9] and set the similarity threshold τ to the noised mean of the non-diagonal elements of the matrix. (2) **Similarity suppression with votes first.** We select synthetic samples in descending order based on their vote counts and add them to the elite set. For each selected sample, we exclude the sample itself and any other samples whose similarity to it exceeds τ in the synthetic sample set by self-similarity matrix S. It drops out the synthetic samples that are highly similar to private ones. (3) **Stepwise selection.** If sample numbers are insufficient, we increase τ by 0.01 and re-execute step (2). This process continues until sufficient samples are selected to form the elite set.

3.4 Elite Set Expansion via Genetic Strategy

To inherit elite samples and generate more diverse samples, we use genetic strategies to expand synthetic data from the elite set, which broadens the search scope during algorithm iterations. Specifically, using the idea of GA, we design three kinds of genetic operations (i.e. mutate, cross, generate) which inherit information from existing elite samples to different degrees to breed new samples via LLMs. Specifically, mutation inherits the information of a single elite sample, crossover inherits the information of multiple samples, and generation does not inherit information, that is, generates completely new samples via metadata and LLMs.

3.4.1 Mutate strategy: editing via LLMs

According to the mutate strategy in GA, which randomly changes the value of the gene locus on sample encoding string, we implement two mutate strategies with text rewriting and text filling.

[7] https://platform.openai.com/docs/guides/embeddings/

[8] Gaussian noise is more suitable for advanced composition (Yu et al., 2024) (see in §4.2), which costs less privacy over iterations, so we add it to the votes in §3.3.1.

[9] Privacy data is usually large, so we randomly sample a portion of it for calculation.

Specifically, for text rewriting, we instruct LLMs to rewrite the given samples from the elite set to achieve sentence structure mutation. For text filling, we aim to achieve semantic information mutation. We first assume text $X = \{x_1, x_2 \ldots, x_n\}$ and employ encoder[10] E to encode text X: $T = E(X) = \{t_1, t_2, \ldots, t_n\}$. Then, we randomly mask some tokens: $T' = \{t_i \text{ if } i \notin M \text{ else } E(\text{'\_'}) \text{ for } i = 1, \ldots, n\}$, where M is randomly selected indices. Next, we employ decoder D to decode T': $X' = D(T') = \{x'_1, \_, \ldots, x'_n\}$. Finally, we prompt LLMs to fill in the blanks '_'. The prompts for these two mutation strategies are shown in App. I.

3.4.2 Cross strategy: sample textual fusion via LLMs

Traditional cross strategy in GA swaps chromosomes of two samples to reorganize the genes for looking forward to better samples, but applying it directly to text will damage the fluency of the text and destroy the context information. We implement the cross strategy on text by in-context learning (ICL). Specifically, we randomly select two samples from the elite set, and then use the content of these two samples to generate new samples via the prompt to achieve the crossover of elite sample information (Specific prompts in App. I).

3.4.3 Generate strategy: creating from metadata via LLMs

Based on the traditional GA idea, we design a new generation strategy to generate completely new samples, which improves the global search ability of the algorithm. Following the design of the initial prompt (see in §3.2), we use privacy metadata such as length distribution to guide LLMs to generate new samples, which can encourage diversity in the evolution of the next generation.

4 Privacy Analysis

We defer DP definitions and lemmas to App. B and only give high-level ideas of the analysis.

4.1 Privacy Analysis for Obtaining Metadata

For the initial prompt design, our privacy is mainly spent on SVT (App. B.2) and attribute histogram noise (length, label, keywords), with budgets of ε_{svt} and ε_{attr} respectively. Since these are macro privacy information and the accuracy requirement is not high, we directly use the Laplace mechanism (App. B.1) to implement $(\varepsilon, 0)$-DP. For SVT implemented using AboveThreshold (see in App. B.2), the sensitivity of its query stream is 1. We add Laplace noise of scale $2/(\varepsilon_{svt}/2)$ and $4/(\varepsilon_{svt}/2)$ to the threshold and query result respectively (the algorithm needs to be executed twice to obtain the minimum and maximum, and the privacy budget is $(\varepsilon_{svt}/2)$ each time). For length and label distribution histograms, the sensitivity is 1. For keyword histograms, we vote according to label division. The privacy data of each label only votes for the keywords corresponding to the label. The sensitivity of this approach is also 1. Finally, we add Laplace noise of scale $1/(\varepsilon_{attr}/3)$ to the three types of histograms respectively.

4.2 Privacy Analysis for Privacy Voting

For privacy voting, we need to use the Analytical Gaussian mechanism (App. B.3) to add Gaussian noise on the voting histograms generated by T iterations, and their L_2 sensitivity is 1, we can use an adaptive combination of Gaussian mechanism sequences to combine these same privacy costs. Specifically, the adaptive composition (App. B.4) of a T identical Gaussian mechanism with a noise multiplier σ satisfies the same privacy guarantee of a single Gaussian mechanism with a noise multiplier $\sigma/\sqrt{T}$. By fixing privacy parameter δ, ε and T, we can calibrate the noise by choosing an appropriate σ in §3.3. In addition, since we perform Poisson subsampling on the privacy data with a sampling rate of q, according to Lemma B.5 (App. B.5), our actual privacy cost is $\varepsilon_{vote} = ln(1 + q(e^{\varepsilon} - 1)$, and $\delta_{vote} = q\delta$.

4.3 Total Privacy Cost

According to the serial composition theorem (App. B.6) for (ε, δ)-DP (Dwork et al., 2006b), the part of initial prompt design satisfies $(\varepsilon_{init}, 0)$-DP, where $\varepsilon_{init} = \varepsilon_{svt} + \varepsilon_{attr}$. The part of Privacy Voting satisfies $(\varepsilon_{vote}, \delta_{vote})$-DP. Therefore, our whole algorithm satisfies $(\varepsilon_{init} + \varepsilon_{vote}, \delta_{vote})$-DP.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate our method DPGA-TextSyn using two datasets. **(1) OpenReview (Xie et al., 2024)** Xie et al. crawl the latest reviews for ICLR 2023 submissions from the OpenReview website to construct this dataset. **(2) PubMed (Yu et al., 2023)** This dataset consists of the abstracts of med-

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

Method	w/o DP	$\varepsilon = 4$	$\varepsilon = 2$	$\varepsilon = 1$
DPSGD+LoRA(8396)	65.1	30.5	30.5	30.5
DPSGD+LoRA(2000)	**55.3**	30.5	30.4	6.3
DP-Transforms(2000)	48.3	38.9	40.4	38.6
AUG-PE(2000)	45.4	43.5	42.8	41.9
Ours (2000)	47.25	**46.14**	**45.21**	**44.53**

Table 1: Results of all baselines on OpenReview. w/o DP indicates no privacy. Numbers in brackets signify the number of training samples, with the first row using all data. The highest accuracy is highlighted in bold.

Method	w/o DP	$\varepsilon = 4$	$\varepsilon = 2$	$\varepsilon = 1$
DPSGD+LoRA(75316)	47.6	34.1	32.5	30.4
DPSGD+LoRA(2000)	**34.6**	1.1	0.8	0.6
DP-Transforms(2000)	33.1	31.2	31.1	31.1
AUG-PE(2000)	32.7	32.5	32.5	32.4
Ours(2000)	33.95	**33.82**	**33.79**	**33.58**

Table 2: Results of all baselines on PubMed.

ical papers from the National Library of Medicine. See the App. C for details of the two datasets.

Baseline. We compare with: **(1) DPSGD+LoRA** (Yu et al., 2021) uses LoRA to fine-tune models under DP. **(2) DP-Transforms** (Yue et al., 2022) generates synthetic data using fine-tuned GPT-2. **(3) AUG-PE** (Xie et al., 2024) uses privacy evolution to guide LLMs to generate synthetic data.

Metrics. We evaluate synthetic data via downstream task accuracy (fine-tuning RoBERTa-base (Liu et al., 2021) for OpenReview text classification, and $\text{BERT}_{\text{Small}}$ (Micheli et al., 2020) for PubMed next-token prediction (Yu et al., 2023)).

Implemention Details and Hyperparameters. We provide them in App. D.

5.2 Overall Performance

Downstream Task Accuracy. Our method outperforms three baselines on OpenReview and PubMed with the same number of training samples for downstream tasks under strong privacy protection (Tab. 1 and Tab. 2). It even exceeds the result of using all privacy data when $\varepsilon = 1, 2$. When without privacy concerns, direct fine-tuning outperforms our method designed for privacy because we just use little privacy information. Our method outperforms the best synthetic data baseline AUG-PE in both privacy and non-privacy situations, and the improvements are significant under the t-test with $p < 0.05$ (details in App. E). Some good synthetic samples are shown in App. J.

5.3 Ablation Studies

	Ours	w/o Length	w/o Keywords	w/o SS
Acc (T=1)	**32.56**	30.12	31.64	31.78
Acc (T=10)	**33.58**	32.08	33.34	32.88

Table 3: Ablation studies on model components. w/o SS stands for w/o Similarity Suppression.

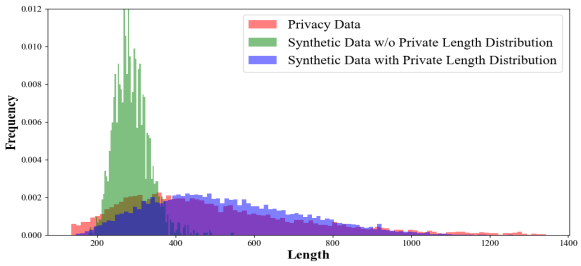

Figure 2: Differences in the length distribution of privacy data and synthetic data for the OpenReview dataset.

As shown in Tab. 3, we conducted an ablation study to evaluate the importance of our model components. Specifically: (1) w/o Length means we do not use the length information of privacy data when synthesizing the initial sample. We can see that without length information, the effect will be significantly reduced. Fig. 2 also shows that the distribution of synthetic data generated by our model without using privacy length information is very different from that of privacy data. (2) w/o Keywords means that we do not use keywords when synthesizing the initial samples, which leads to performance degradation, indicating the importance of descriptive metadata for sample synthesis. (3) w/o Similarity Suppression means that when we select elite samples, we only select samples with high votes without considering their similarity. Fig. 3 shows that if the similarity is not suppressed when executing our method, the self-similarity of synthetic samples will gradually increase with iterations (specific samples can be viewed in App. H). These result in reduced effects, indicating that similarity suppression is essential.

5.4 Additional Analysis of Our Method

Distribution Similarity to Privacy Data. We evaluate the relevance of synthetic and privacy data via MAUVE (Pillutla et al., 2021) and FID (Heusel et al., 2017) and record these two metrics changes as our method iterates. Fig. 4 shows the MAUVE

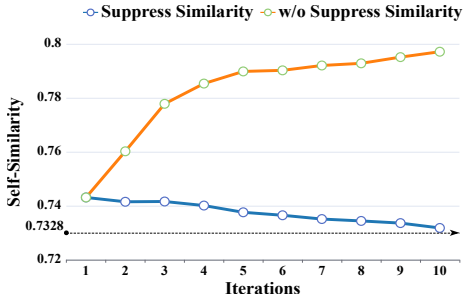

Figure 3: Changes in self-similarity metric during iterations, with two curves representing whether to suppress similarity when selecting elite samples.

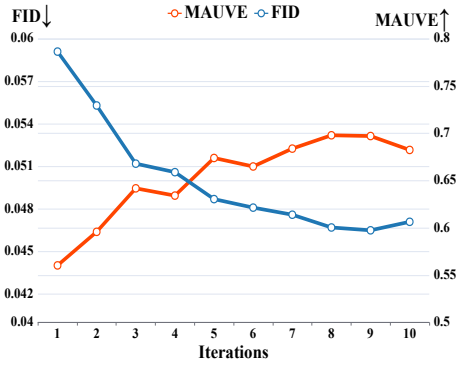

Figure 4: MAUVE and FID metrics.

and FID metrics gradually get better over algorithm iterations, indicating that our synthetic data is becoming increasingly similar to privacy data, thus demonstrating the effectiveness of our method.

Self-Similarity. We use the self-similarity metric to measure the diversity of synthetic samples from our model. The blue line in Fig. 3 shows that the self-similarity of our model slowly decreases with iterations, approaching the self-similarity of privacy data[11]. This shows that our method effectively improves the diversity of synthetic data and avoids samples that are too similar.

We also conducted inference attack experiments to verify the privacy protection capabilities of our method. For details, please refer to App. F.

5.5 Research on Keywords Selection

To verify the role of keyword selection, we synthesize the initial samples without and with selecting keywords. We use t-SNE to reduce the dimension of the embeddings of initial synthetic data generated by the above two ways and privacy data, and then we visualize them. It is shown in Fig. 5, where blue points on the left and right are for synthetic data without and with selecting keywords, and green points are for privacy data. We can see that the left figure shows a large difference between the distribution of the synthetic samples generated by unused keyword selection and the privacy data, but the right figure indicates that the initial synthetic data and privacy data are closer in distribution. Therefore, it is important that we select the keywords by privacy voting, which can avoid generating many irrelevant samples.

[11]The diversity of synthetic data gradually remains consistent with private data with iterations.

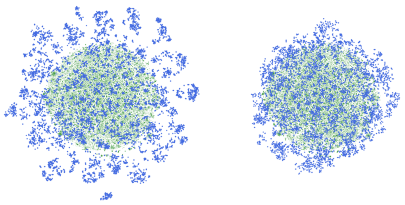

Figure 5: The t-SNE visualization of data embeddings.

5.6 Analysis on Similarity Threshold

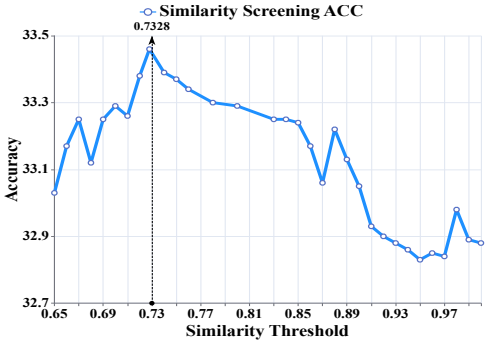

Figure 6: Impact of similarity threshold changes on accuracy.

For the similarity threshold, we conducted experiments for our model on PubMed to explore how different similarity thresholds affect the accuracy of synthetic data in downstream tasks. The results are displayed in Fig. 6. We found that both too high and too low similarity thresholds result in poor performance. It only shows the best performance when it is close to the self-similarity of the privacy data, which can also be seen from Fig. 3.

6 Conclusion

In summary, we propose DPGA-TextSyn, a method for privacy text synthesis under DP based on genetic algorithm and LLMs. First, we integrate privacy metadata information into the initialization prompt to generate better initial synthetic samples.

Then, we construct a DP histogram based on nearest neighbor voting and a self-similarity matrix to select distinct elite samples. Finally, we use three different genetic strategies including mutation, crossover, and generation to expand the elite set. Experiments show that our method excels in text synthesis, outperforming all baseline methods.

7 Limitations

Our work proposes DPGA-TextSyn to produce high-utility synthetic text under DP constraints, but two key limitations warrant discussion.

First, we used GPT-3.5 in our experiments, not the latest and SOTA LLMs like OenpAI's GPT-4 or o1 models, which seems to limit the performance of our model. The main reason is that our baseline AUG-PE is based on GPT-3.5 for fair comparison. In addition, compared with GPT-3.5, it is too expensive to obtain the API key of GPT-4 or o1 models. The instruction-following ability of GPT-3.5 is not as good as the most advanced LLMs at present, and it is more sensitive to the quality of prompt design. In the future, we will try to use more advanced models for effect verification.

This paper aims to mitigate the risk of failing to balance the quality of private and synthetic data when leveraging large language models. However, our method does not provide perfect protection. We recommend that users employ the process with caution, avoiding complete reliance on the tool to prevent potential privacy leakage and address ethical considerations.

8 Ethical Considerations

We place significant importance on ethical considerations and adhere rigorously to the ACL Ethics Policy.

(1) Our study proposes a novel method to synthesize text under differential privacy via LLMs, which does not require consideration of ethical issues regarding motivation or algorithmic approaches, as no private information is used.

(2) In general, if the dataset contains privacy information, it may be leaked during use. All the data sets used in our experiments have been published, allowing us to conduct experiments to evaluate the effectiveness of privacy protection measures.

(3) Moreover, it's imperative to exercise caution in utilizing our model and refrain from assuming its infallibility. One potential unethical application involves gathering data from users who believe our model guarantees complete privacy protection, potentially overlooking the actual strength of privacy safeguards. This oversight could lead to adverse societal consequences.

Acknowledgements

This work is supported by the following foundations: the National Natural Science Foundation of China under Grant No.62025208 and 62421002, the National Science Foundation for Distinguished Young Scholars under Grant No.62325604, No.62125604, the Young Elite Scientist Sponsorship Program by CAST under Grant No.YESS20230367, and the National Natural Science Foundation of China (NSFC) under Grant No.62306330, No.62376284.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*.

Borja Balle, Gilles Barthe, and Marco Gaboardi. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31.

Borja Balle and Yu-Xiang Wang. 2018. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR.

Maurice Stevenson Bartlett. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. 2010. *Bounds on the Sample Complexity for Private Learning and Private Data Release*. Theory of Cryptography.

Johannes Breuer, Libby Bishop, and Katharina Kinder-Kurlanda. 2020. The practical and ethical challenges

in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, 22(11):2058–2080.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Aldo Gael Carranza, Rezsa Farahani, Natalia Ponomareva, Alex Kurakin, Matthew Jagielski, and Milad Nasr. 2023. Synthetic query generation for privacy-preserving deep retrieval systems using differentially private language models. *arXiv preprint arXiv:2305.05973*.

TCG CREST. 2023. A survey on privacy preserving synthetic data generation and a discussion on a privacy-utility trade-off problem. In *Science of Cyber Security-SciSec 2022 Workshops: AI-CryptoSec, TA-BC-NFT, and MathSci-Qsafe 2022, Matsue, Japan, August 10–12, 2022, Revised Selected Papers*, page 167. Springer Nature.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. 2022. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*.

Emiliano De Cristofaro. 2023. What is synthetic data? the good, the bad, and the ugly. *arXiv preprint arXiv:2303.01230*, page 60.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. 2022. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*.

Jinshuo Dong, Aaron Roth, and Weijie J Su. 2022. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37.

C. Dwork, K. Kenthapadi, Frank Mcsherry, Ilya Mironov, and M. Naor. 2006a. Our data, ourselves: Privacy via distributed noise generation. *Springer, Berlin, Heidelberg*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Cynthia Dwork. 2011. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

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

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Fredrik Harder, Milad Jalali Asadabadi, Danica J Sutherland, and Mijung Park. 2022. Pre-trained perceptual features improve differentially private image generation. *arXiv preprint arXiv:2205.12900*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. 2014. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

Peter Kairouz, Ziyu Liu, and Thomas Steinke. 2021. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, pages 5201–5212. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin Zhang. 2022. Hard gate knowledge distillation - leverage calibration for robust and

reliable language model. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9793–9803, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaewoo Lee and Daniel Kifer. 2018. Concentrated differentially private gradient descent with adaptive periteration privacy budget. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1656–1665.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. 2023. Differentially private synthetic data via foundation model apis 1: Images. *arXiv preprint arXiv:2305.15560*.

Zinan Lin, Vyas Sekar, and Giulia Fanti. 2021. On the privacy properties of gan-generated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Min Lyu, Dong Su, and Ninghui Li. 2016. Understanding the sparse vector technique for differential privacy. *arXiv*.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. *arXiv preprint arXiv:2210.13918*.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. 2022. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *arXiv preprint arXiv:2201.12677*.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. *arXiv preprint arXiv:2010.03813*.

Pablo A Osorio-Marulanda, John Esteban Castro Ramirez, Mikel Hernández Jiménez, Nicolas Moreno Reyes, and Gorka Epelde Unanue. 2024. Differentially private non parametric copulas: Generating synthetic data with non parametric copulas under privacy guarantees. *arXiv preprint arXiv:2409.18611*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Pranav Putta, Ander Steele, and Joseph W Ferrara. 2022. Differentially private conditional text generation for synthetic data production.

Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. 2020. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv preprint arXiv:2011.05537*.

Paul Schmiedmayer, Adrit Rao, Philipp Zagar, Vishnu Ravi, Aydin Zahedivash, Arash Fereydooni, and Oliver Aalami. 2024. Llm on fhir–demystifying health records. *arXiv preprint arXiv:2402.01711*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2023. Privacy-preserving in-context learning with differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*.

Zhiliang Tian, Yingxiu Zhao, Ziyue Huang, Yu-Xiang Wang, Nevin L. Zhang, and He He. 2022. Seqpate: Differentially private text generation via knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 35, pages 11117–11130. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. 2024. Differentially private synthetic data via foundation model apis 2: Text. *arXiv preprint arXiv:2403.01749*.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.

Yucheng Yin, Zinan Lin, Minhao Jin, Giulia Fanti, and Vyas Sekar. 2022. Practical gan-based synthetic ip

header trace generation using netshare. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 458–472.

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. 2023. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine*, 6(1):141.

Da Yu, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Janardhan Kulkarni, Zinan Lin, Chulin Xie, Huishuai Zhang, and Wanrong Zhang. 2023. Training private and efficient language models with synthetic data from llms. In *Socially Responsible Language Modelling Research*.

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. 2024. Privacy-preserving instructions for aligning large language models. *arXiv preprint arXiv:2402.13659*.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2022. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*.

Yanqing Zhang, Qi Xu, Niansheng Tang, and Annie Qu. 2024. Differentially private data release for mixed-type data via latent factor models. *Journal of Machine Learning Research*, 25(116):1–37.

Haoqi Zheng, Qihuang Zhong, Liang Ding, Zhiliang Tian, Xin Niu, Changjian Wang, Dongsheng Li, and Dacheng Tao. 2023. Self-evolution learning for mixup: Enhance data augmentation on few-shot text classification tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8964–8974, Singapore. Association for Computational Linguistics.

Kangchen Zhu, Zhiliang Tian, Jingyu Wei, Ruifeng Luo, Yiping Song, and Xiaoguang Mao. 2024. StyleFlow: Disentangle latent representations via normalizing flow for unsupervised text style transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15384–15397, Torino, Italia. ELRA and ICCL.

Yuqing Zhu and Yu Xiang Wang. 2020. Improving sparse vector technique with renyi differential privacy. In *Neural Information Processing Systems*.

A Additional Related Work

Privacy Synthetic Data

Synthetic data generation is a critical strategy in the field of data privacy protection (Jordon et al., 2022). It can generate datasets that mimic real data's statistical properties while removing personally identifiable information (Yoon et al., 2023), thereby safeguarding individual privacy (Osorio-Marulanda et al., 2024). Liu et al. proposed an integrated method to combine different techniques, improving diversity and utility (Kairouz et al., 2021). The AIM method uses a select-measure-generate approach, enhancing the selection phase by greedily choosing the most informative queries to generate synthetic data (McKenna et al., 2022). For complex data types like text, images, and video, most methods use generative models, such as Generative Adversarial Networks (GANs) (Xie et al., 2018). PATE-GAN explores GAN-based methods (Jordon et al., 2018) to produce synthetic data through adversarial training. However, traditional synthetic data often lacks standardized privacy protection strategies in the de-identification process (CREST, 2023). Thus, conventional synthetic data has significant privacy limitations, making strong privacy guarantees challenging (De Cristofaro, 2023).

DP synthetic data surpasses traditional synthetic data in privacy protection (Rosenblatt et al., 2020). DP generates synthetic data resembling the original while obscuring individual data points, preventing adversary identification (Lin et al., 2021; Dwork et al., 2006a; Abadi et al., 2016). This is achieved by incorporating randomness to reduce individual data points' influence on the dataset (Tang et al., 2023). DP-SGD ensures differential privacy by bounding gradient sensitivity and injecting Gaussian noise during training (Abadi et al., 2016). Dockhorn et al. extended DP-SGD to train Generative Adversarial Networks and diffusion models (Dockhorn et al., 2022). Yin and Yu demonstrated that pre-training on public data and DP-SGD fine-tuning improve generative models' privacy-utility trade-offs (Yue et al., 2022; Yin et al., 2022). DP-MEPF trains generative models to generate synthetic data that preserves private features' statistics while ensuring privacy (Harder et al., 2022). For deep retrieval systems, DP language models generate synthetic queries resembling original data, enabling secure training of deep retrieval systems while ensuring privacy (Kurakin et al., 2023; Carranza et al., 2023).

In addition, some work has attempted to use distillation techniques in the context of privacy (Tian et al., 2022; Lee et al., 2022). In the future, transfer learning (Zhu et al., 2024) and data augmentation (Zheng et al., 2023) can also be explored in combination with privacy.

B Differential Privacy

B.1 The Laplace Mechanism

Definition B.1 (The Laplace Mechanism(Dwork et al., 2014))). Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as: $\mathcal{M}_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \ldots, Y_k)$, where Y_i are i.i.d. random variables drawn from $\text{Lap}(\Delta/\varepsilon)$. The Laplace mechanism preserves $(\varepsilon, 0)$-DP.

B.2 Sparse Vector Technique (SVT)

Definition B.2 (Sparse Vector Technique(Zhu and Wang, 2020; Lyu et al., 2016)). The Sparse Vector Technique (SVT) is designed to identify queries whose results exceed a certain threshold while maintaining privacy. In the SVT framework, the input comprises a sequence of queries $q_1, q_2, \ldots, q_i, \ldots \in \mathcal{Q}(\Delta)$ along with a sequence of thresholds $T_1, T_2, \ldots, T_k, \ldots$. The algorithm aims to generate a binary vector $\{\perp, \top\}$, where $\top$ indicates the query result exceeds the threshold, and $\perp$ indicates it is below. See Alg. 1 for the algorithm process.

Algorithm 1 Input is a privacy database D, an adaptively chosen stream of sensitivity 1 queries $f_1, \ldots$, and a threshold T. Output is a stream of responses $a_1, \ldots$

AboveThreshold $(D, \{f_i\}, T, \varepsilon$
Let $\hat{T} = T + Lap\left(\frac{2}{\varepsilon}\right)$.
for Each query i **do**
    Let $v_i = Lap\left(\frac{4}{\varepsilon}\right)$
    if $f_i(D) + v_i \geq \hat{T}$ **then**
        Output $a_i = \top$.
        Halt.
    else
        Output $a_i = \perp$.
    end if
end for

B.3 Analytical Gaussian mechanism

Lemma B.3 (Analytical Gaussian mechanism (Balle and Wang, 2018)). For a numeric query $f : X^n \to \mathbb{R}^d$ over a dataset $\mathcal{D}$, the randomized algorithm that outputs $f(\mathcal{D}) + Z$ where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$ satisfies $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geq 0$ and $\delta(\varepsilon) = \Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right)$. Here, $\Delta := \Delta_2^{(f)} = \max_{\mathcal{D}\sim\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2$ is the global L_2 sensitivity of f and Φ is the CDF function of $\mathcal{N}(0, 1)$.

B.4 Composition of Gaussian mechanisms

Lemma B.4 (Composition of Gaussian mechanisms (Dong et al., 2022)). The adaptive composition of a sequence of Gaussian mechanisms with a noise level $\sigma_1, \sigma_2, \ldots$ and global L_2 sensitivity $\Delta_1, \Delta_2, \ldots$ satisfies $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geq 0$ and $\delta(\varepsilon) \leq \delta_{\mathcal{M}}(\varepsilon)$ where $\mathcal{M}$ is a Gaussian mechanism with noise multiplier $\sigma/\Delta = \left(\sum_i (\Delta_i/\sigma_i)^2\right)^{-1/2}$.

B.5 Privacy Amplification by Subsampling

Lemma B.5 (Privacy Amplification by Subsampling(Balle et al., 2018; Abadi et al., 2016)). Privacy Amplification by Subsampling is a technique that enhances privacy by operating on a randomly selected subset of the dataset (Beimel et al., 2010). In this process, a mechanism that satisfies (ε, δ)-DP for a dataset is applied to a subset sampled independently with a fixed probability q. The result is that the algorithm achieves stronger privacy guarantees, typically $(ln(1 + q(e^{\varepsilon} - 1), q\delta)$-DP for the entire dataset.

B.6 Serial Composition of Privacy

Lemma B.6 (Serial Composition of Privacy(Dwork et al., 2014)). Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_i$ be an $(\varepsilon_i, \delta_i)$-DP for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \to \prod_{i=1}^k \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), \ldots, \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i)$-DP.

C Details of Datasets

OpenReview Dataset: We utilized the dataset compiled by Xie et al., which was scraped from the OpenReview website using the overview-py library, specifically focusing on ICLR 2023 data. In this dataset, the fields of paper abstracts and reviews were merged into a single sample. The dataset was categorized into 12 classes based on the research domain of the papers. Classes with fewer than 100 samples in the training set were removed. The training, validation, and test datasets consist of 8,396, 2,798, and 2,798 samples, respectively.

PubMed Dataset: We utilized the dataset compiled by Yu et al., which contains abstracts of medical papers scraped from the U.S. National Library of Medicine. These abstracts were published between 2023/08/01 and 2023/08/07. The training set consists of 75,329 abstracts from 08/01 to 08/05. The validation set includes 4,453 abstracts from 08/06, while the test set comprises 14,423 abstracts from 08/07.

D Implemention Details and Hyperparameters

D.1 Model

We use GPT-4 (Achiam et al., 2023) to generate keywords (subcategories for OpenReview, medical professions for PubMed) with a temperature of 1.4 for diversity. For synthetic sample generation, we use prompt (see in APP. I) to guide GPT-3.5 to generate initial samples and implement mutation, crossover, and genetics to expand samples, the temperature is set to 1.2. For the embedding model, we use text-embedding-ada-002 from OpenAI.

D.2 Downstream Tasks

Following previous work (Xie et al., 2024), for OpenReview, we finetune the RoBERTa-base (Liu et al., 2021) model for text classification tasks (the label is Area). 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 10. For PubMed, We fine-tune $\text{BERT}_{\textbf{Small}}$ (Micheli et al., 2020) for the next token prediction task. 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, and the weight decay as 0.01. We finetune 10 epochs. We report the mean of our method three runs for all privacy budgets.

D.3 Privacy Settings

Following previous work (Yu et al., 2021; Yue et al., 2022; Xie et al., 2024), we set the overall privacy parameters $\varepsilon = 1, 2, 4, \infty$, and $\delta = 1/(N \cdot \log N)$, which should be smaller than the inverse of the dataset size N (Hsu et al., 2014; De et al., 2022). For different privacy budgets, we always allocate 0.5 of the privacy budget to the acquisition of privacy metadata, 0.2 of which is used to perform SVT, 0.3 is used to add noise to three different histograms, and the rest is used for privacy voting when selecting the elite set.

D.4 Other Hyperparameters

When constructing the length histogram, we set the number of bins to 100. When voting, we set the subsampling rate of privacy data to 0.8. When performing diversity suppression, we increase the diversity by 0.01 each time. When performing a text fill mutant operation, we set the probability that each token is masked to 0.5. For both tasks, we set the number of initial synthetic data to 10,000, the number of selected elite sets to 2,000, and then use two genetic operations to expand 4,000 samples, use the crossover operation to expand 2,000 samples, and use the generation operation to generate 2,000 new samples. These expanded samples, together with the original elite set, form the next generation of initial 10,000 samples.

E Significance Test Results

We conduct the t-test (Bartlett, 1937) to examine whether the improvements of our method are significant. The p values in Tab. 4 are all smaller than 0.05, demonstrating the significance of our improvements.

F Inference Attack Experiment

We used synthetic Openreview data to fine-tune BERT for membership inference attack (Shokri et al., 2017) experiments, the result is shown in Tab. 5. We found that the attack success rate of the model fine-tuned with synthetic data was lower than that of the model fine-tuned directly with privacy data, indicating that our design can effectively resist privacy inference attacks.

G Research on Subsampling Rate

We also researched the sampling rate during the subsampling of privacy data in the PubMed dataset. We found that setting an appropriate subsampling rate can improve synthetic data's effectiveness in downstream tasks to a certain extent. We believe that although subsampling privacy data reduces some privacy information, it can involve adding less noise under a fixed privacy budget. When a certain balance is achieved, optimal results can be realized. Fig. 7 displays the results of our experiments. Under the condition that the privacy budget ε is set to 1, the highest accuracy of the synthetic samples was achieved when the subsampling rate was 0.8, which was even higher than that of synthetic samples without subsampling.

Datasets	Openreview			PubMed		
Privacy Budget	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$
p	9.04e-7	7.84e-6	2.45e-5	1.33e-4	7.79e-5	1.35e-4

Table 4: The p values of t-test on our method with baseline AUG-PE. The p values are all smaller than 0.05, indicating our improvements are significant.

Method	Entropy	Confidence
w/o DP	0.568	0.561
Ours($\varepsilon = 1$)	0.503	0.496

Table 5: Method and corresponding entropy and confidence.

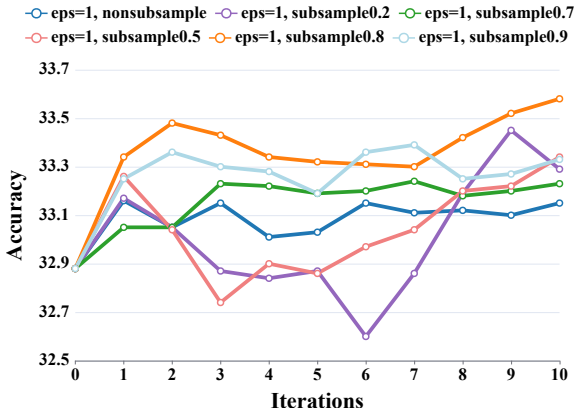

Figure 7: Accuracy changes with iterations under different subsampling rates.

H Examples of High Similarity Samples

Some high-similarity samples appear in the final synthesized data when similarity is not suppressed during the selection of elite offspring. The results are shown in Tab. 6

I APIs Prompt Designs

The specific content of the prompt for each API of our method can be found in Tab. 7 and Tab. 9. Some keywords generated for each dataset using GPT-4 can be found in Tab. 11 and Tab. 8, and the specific prompts generated are shown in Tab. 10

J Synthetic Data Examples

Some examples of OpenReview and PubMed datasets generated using our method are shown in Tab. 12 and Tab. 13.

Example 1	Example 2	Similarity
Abstract: This study aimed to investigate the potential association between the ongoing COVID-19 pandemic and the pathogenesis of neurological complications. Our research involved a comprehensive analysis of clinical and experimental data to evaluate the impact of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) on the central nervous system (CNS). Through literature review and case studies, we provide compelling evidence of the presence of SARS-CoV-2 in the CNS and its ability to induce a range of neurological manifestations. Additionally, we explore potential mechanisms underlying SARS-CoV-2 neuroinvasion and discuss the potential long-term impacts on neurologic function.	Abstract: This study investigates the potential relationship between the CoViD-19 pandemic and the pathology of neurological complications. Our research utilized comprehensive clinical and patient data to evaluate the impact of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) on the central nervous system(CNS). Through in-depth analysis and examination, we present compelling evidence regarding the presence of SARS-CoV-2 in the CNS and its ability to induce a diverse range of neurological manifestationsFurthermore, we explore potential mechanisms underlying SARS-CoV-2neuroinvasion and the potential long-term implications on neurological health.	0.97
An investigation was conducted to evaluate the efficacy of direct oral anticoagulants (D0ACs) compared to warfarin in patients with non-valvular atrial fibrillation (NVAF). In this study spanning from January 2015 to December 2019, there was a significant preference for DOACsover warfarin due to lower rates of adverse events in patients with AF. These findings indicate the potential for DOACs to improve outcomes in the NVAF population, underscoring the importance of selecting appropriate therapy for optimal patient outcomes. Further research on varying patient groups and different anticoagulants is warranted to enhance our understanding of stroke prevention in patients The utilization of real-world data in the study adds a layer of insights that can contribute to the prevention of adverse events and the advancement of anticoagulant therapy. By making treatment decisions based on individual patient profiles, healthcare providers can significantly improve the quality of life and outcomes of patients with NAF. Considering a personalized perspective on individual profiles can lead to better outcomes for patients, emphasizing the importance of exploring novel prevention strategies to enhance the well-being of patients with AF and improve prevention outcomes.	An elegant study was conducted to evaluate the efficacy and safety of direct oral anticoagulants (DOACs) compared to warfarin in patients with non-valvular atrial fibrillation (NVAF). Over a thorough investigation spanning from January 2015 to June 2019, there was a significant preference for DOACs over warfarin due to lower rates of adverse events in patients with NVAF, These findings indicate the potential role of DOACs in improving outcomes in the NVAF population, underscoring the importance of selecting appropriate anticoagulants for optimal patient outcomes, Further research on varying patient groups and different anticoagulants is crucial to expanding our understanding of stroke prevention in patients with NVAF The real. world data in this study adds another layer of understanding to the occurrence of events and the advancement of anticoagulation therapy. By making personalized decisions based on individual patient characteristics, healthcare professionals can effectively improve the quality and outcomes of patients with NVAF. Personalized and individualized care for patients underscores the importance of implementing novel prevention strategies to enhance the well-being and prevention outcomes of NVAF patients.	0.96

Table 6: Some high-similarity samples appear in the final synthesized data when similarity is not suppressed.

Operators	Prompts
initialize	"system": "Given the area of the research paper, you will need to provide a sample review based on the area subcategory, recommendation, and word count requirements provided, including the following: 1. summarizing the paper in detail; 2. listing the strengths and weaknesses of the paper in detail; 3. summarizing the review in detail." "user": "{label of area} | Subcategory Area: {random_sample_subcategory} | {random_sample_recommendation}. The generated review must have {random_sample_length} words."
mutate1	"system": "You are a helpful assistant for text rewriting." "user": "{old sample} The above is a review from the OpenReview website. Please rewrite this review from a professional perspective in a different style and tone, but ensure that the area and recommendation do not change. The generated new sample must have {the length of the old sample} words, please directly generate the new review without other unnecessary content or reminders."
mutate2	"system": "You are a helpful assistant for text filling." "user": "{old sample} The above is a review from the OpenReview website. Please refer to the style of review in OpenReview, guess the content of the blank "_", and fill in the blank to generate a new review, but ensure that the area and recommendation do not change. The generated new sample must have {the length of the old sample} words, please directly generate the new review without other unnecessary content or reminders."
cross	"system": "You are a helpful assistant." "user": "{old samples} The above are reviews on the OpenReview website. Please refer to the style of the sample reviews provided above, and the area subcategory and recommendation provided below to generate a new review. {label of area} | Subcategory Area: {random_sample_subcategory} | {random_sample_recommendation}. Please make sure that the newly generated review is not too similar to the original review, and the new review must have {the average length of the old samples} words. Please generate a new review directly without adding other unnecessary content or reminders."
generate	"system": "Given the area of the research paper, you will need to provide a sample review based on the area subcategory, recommendation, and word count requirements provided, including the following: 1. summarizing the paper in detail; 2. listing the strengths and weaknesses of the paper in detail; 3. summarizing the review in detail." "user": "{label of area} | Subcategory Area: {random_sample_subcategory} | {random_sample_recommendation}. The generated review must have {random_sample_length} words."

Table 7: Prompts for different types of APIs when synthesizing data using the OpenReview dataset

Medical Careers
Medical Doctor (MD), Surgeon, Pharmacist, Nurse Practitioner, Physician Assistant, Public Health Researcher, Epidemiologist, Biostatistician, Molecular Biologist, Geneticist, Neuroscientist, Pathologist, Immunologist, Microbiologist, Bioinformatician, Health Policy Expert, Toxicologist, Dentist, Veterinary Scientist, Physiotherapist, Nutritionist, Dietitian, Healthcare Administrator, Social Worker, Psychiatrist, Radiologist, Cardiologist, Endocrinologist, Pediatrician, Oncologist, Dermatologist, Orthopedist, Ophthalmologist, Gynecologist, Anesthesiologist, Rheumatologist, Virologist, Nephrologist, Clinical Researcher, Principal Investigator, Biomedical Engineer, Psychologist

Table 8: Medical careers keywords generated for the PubMed dataset

Operators	Prompts
initialize	"system": "You are a useful assistant for generating abstracts of medical papers." "user": "Generate an abstract of a medical research paper for {random_sample_author}, imitating the standard, format, and style of PubMed journal articles. Please ensure that the abstract is professional and appropriate for a scientific journal, and utilize diverse sentence structures and advanced grammatical constructs to enhance readability. The generated abstract must have {random_sample_length} words, No more, no less. No title and other content are needed, please directly generate the abstract content without the word "abstract"."
mutate1	"system": "You are a helpful assistant for text rewriting." "user": "{old sample} Above is an abstract of a medical paper from PubMed journal, please refer to the style of PubMed medical paper abstracts, and rewrite this abstract provided above to generate a new abstract by modifying its experimental methods, experimental data, research topic, research subjects, styles, tones, and other related content. Require the newly generated abstract to have a greater degree of rewriting compared to the original abstract, and not to be too similar to the original abstract. Please ensure that the new abstract is professional and appropriate for a scientific journal, and utilize diverse sentence structures and advanced grammatical constructs to enhance readability. The generated abstract must have {the length of the old sample} words, No more, no less. No title and other content are needed, please directly generate the new abstract content without the word "abstract" and other unnecessary content or reminders."
mutate2	"system": "You are a helpful assistant for text filling." "user": "{old sample with blank} Above is an abstract of a medical paper with blank spaces from PubMed journal, please refer to the style of PubMed medical paper abstracts, guess the content of the blank "_", and fill in the blank to generate a new abstract using different styles and tones. If there are no blanks, please output the original medical abstract. Please ensure that the new abstract is professional and appropriate for a scientific journal, and utilize diverse sentence structures and advanced grammatical constructs to enhance readability. The generated abstract must have {the length of the old sample} words, No more, no less. No title and other content are needed, please directly generate the new abstract content without the word "abstract" and other unnecessary content or reminders."
cross	"system": "You are a helpful assistant." "user": "{old samples} Above are abstracts of medical papers from PubMed journal, please refer to the style of PubMed medical paper abstracts, and imitate the format and related content of the medical paper abstracts provided above to generate a new abstract. Please ensure that the new abstract is professional and appropriate for a scientific journal, and utilize diverse sentence structures and advanced grammatical constructs to enhance readability. The generated abstract must have {the average length of the old samples} words, No more, no less. No title and other content are needed, please directly generate the new abstract content without the word "abstract" and other unnecessary content or reminders."
generate	"system": "You are a helpful assistant." "user": "Please refer to the style of PubMed medical paper abstracts to generate a new abstract for {random_sample_author}. Please ensure that the new abstract is professional and appropriate for a scientific journal, and utilize diverse sentence structures and advanced grammatical constructs to enhance readability. The generated abstract must have {random_sample_length} words, No more, no less. No title and other content are needed, please directly generate the new abstract content without the word "abstract" and other unnecessary content or reminders."

Table 9: Prompts for different types of APIs when synthesizing data using the PubMed dataset

Datasets	Prompts
OpenReview	"system": "You are a useful assistant." "user": "{Label information for OpenReview}, I want to break down the main categories of ICLR23. The above are some main categories. Please provide me with a list of possible subcategories for each major category, and try not to duplicate subcategories for different major categories. Each major category should have at least 50 subcategories. Output in JSON format, JSON format does not require numerical sequence numbers, only "Area" and "Subcategories" are needed."
PubMed	"system": "You are a useful assistant." "user": "Please refer to the PubMed website to generate 50 different medical professions for me, without too much duplication between them, and output them in json format."

Table 10: Prompts for getting keywords using GPT-4

Labels	Keywords
Deep Learning and representational learning	"Convolutional neural networks (CNNs)", "Recurrent neural networks (RNNs)", "Long short-term memory networks (LSTMs)", "Autoencoders", "Deep belief networks", "Generative adversarial networks (GANs)", "Transformer architectures", "Capsule networks", "Deep reinforcement learning", "Attention mechanisms", "Feature learning", "Deep learning optimization techniques", "Neural architecture search", "Meta-learning in deep architectures", "Transfer learning", "Multi-task learning", "End-to-end learning", "Layer-wise training techniques", "Representation learning for text", "Image representation learning", "Audio and speech representation learning", "Multimodal learning", "Graph neural networks", "Deep learning for structured data", "Exploration of learning rates", "Regularization techniques", "Early stopping", "Dropout techniques", "Batch normalization", "Depth and width of networks", "Energy-efficient deep learning", "Hardware accelerations for deep learning", "Deep learning compilers", "Deep learning for embedded systems", "Quantum deep learning", "Deep learning in edge devices", "Scalability in deep learning", "Robustness in deep models", "Benchmarking deep learning frameworks", "Deep learning in adverse conditions", "Bias and fairness in deep learning", "Compression of deep models", "Interpretability of deep representations", "Adversarial examples in deep learning", "Privacy-preserving deep learning", "Federated deep learning", "Continual learning", "Synthetic data for deep learning", "Self-organized deep learning"
Applications (eg, speech processing, computer vision, NLP)	"Speech recognition", "Speech synthesis", "Speech enhancement", "Natural language understanding", "Natural language generation", "Machine translation", "Semantic analysis", "Sentiment analysis", "Computer vision for medical diagnostics", "Self-driving car vision systems", "Augmented reality", "Virtual reality", "Facial recognition", "Image classification", "Object detection", "Semantic segmentation", "Instance segmentation", "Pose estimation", "Optical character recognition (OCR)", "Video analytics", "Surveillance systems", "Remote sensing", "Drone vision", "Robotics vision systems", "Agricultural monitoring", "Photogrammetry", "Multimedia systems", "Interactive systems", "Automated customer support", "Predictive maintenance", "Supply chain automation", "Fraud detection", "Recommendation systems", "Advertising systems", "Personalization technologies", "E-commerce systems", "Content moderation", "Human-robot interaction", "Accessible technologies", "EdTech", "HealthTech", "FinTech", "Energy sector applications", "Climate monitoring systems", "Precision agriculture",

Table 11: Keywords generated for the OpenReview dataset example labels

Examples	Labels
Summary: The paper focuses on deep learning and representation learning techniques, aiming to propose a new algorithm to enhance model performance. The authors provide a thorough explanation of the proposed method and conduct experiments to compare it with existing approaches. Results indicate a marginal improvement in performance compared to baseline models. Strengths: 1. Clear explanation of the proposed algorithm. 2. Rigorous experimental methodology. 3. Marginal improvement in model performance. Weaknesses: 1. Lack of comparison with state-of-the-art methods. 2. Limited discussion on the algorithm's computational complexity. 3. The paper could benefit from more real-world applications or case studies. Overall, the paper presents a well-defined approach to enhancing deep learning models. The authors make a solid effort to conduct comprehensive experiments, although further improvements are necessary to better position the proposed algorithm among state-of-the-art methods. Considering the marginal performance improvement, the paper is recommended for publication, slightly above the acceptance threshold.	Deep Learning and representational learning
1. Summary: The research paper proposed a novel approach to enhancing image classification accuracy using variational inference techniques. The authors demonstrated the effectiveness of their method through extensive experiments on benchmark datasets, achieving a marginal improvement in classification performance compared to existing methods. The paper provided a thorough analysis of the proposed approach, including the mathematical formulation of the variational inference framework and its implementation in the context of image classification tasks. **Strengths:** - The paper addresses an important problem in the field of image classification and provides a novel solution using variational inference techniques. - The experimental results demonstrate the efficacy of the proposed approach, showing a marginal improvement in classification accuracy. - The mathematical derivations and implementation details are presented, making it easy for readers to understand and replicate the methodology. **Weaknesses:** - The paper could benefit from a more comprehensive discussion of the limitations of the proposed approach and potential directions for future research. - The experimental evaluation could be expanded to include more diverse datasets to further validate the generalizability of the proposed method. - The significance of the marginal improvement in classification accuracy compared to existing methods could be better emphasized in the paper. **3. Review Summary:** Overall, the research paper makes a valuable contribution to the field of image classification by introducing a novel approach based on variational inference techniques. While the paper is well-written and the methodology is sound, addressing the identified weaknesses would further strengthen the paper and justify its recommendation for publication marginally above the acceptance threshold.	Probabilistic Methods (eg, variational inference, causal inference, Gaussian processes)

Table 12: Some examples of synthetic samples generated using the OpenReview dataset

Examples
Introduction: Nurse practitioners (NPs) play a vital role in providing patient-centered care, especially in rural areas with limited access to primary care physicians. Given their increasing prominence, the literature has recognized the importance of evaluating the educational preparedness of NPs to meet the demands of their role effectively. This quantitative study aims to examine the relationship between NP's level of education and patient outcomes. By analyzing a nationwide sample of 1,000 NPs, the study found a statistically significant positive correlation between higher levels of education (master's and doctoral degrees) and improved patient outcomes across various health indicators. These findings support the need for continued investment in advanced NP education to optimize patient outcomes.
Pelvic inflammatory disease (PID) is a common cause of morbidity among women of reproductive age, with substantial healthcare costs and long-term sequelae. The primary goal of this study was to evaluate the efficacy of prophylactic antibiotic treatment in preventing subsequent episodes of PID in high-risk women. A randomized, double-blind, placebo-controlled trial was conducted from January 20XX to December 20XX, recruiting women aged 18-45 years who presented with symptoms consistent with acute PID. Participants were randomly assigned to receive either a 14-day course of broad-spectrum antibiotics or a placebo. They were followed up for a period of 12 months, during which they were regularly assessed for clinical symptoms, evaluated for treatment compliance, and underwent laboratory investigations including pelvic ultrasound and microbiological testing. Our results demonstrated that prophylactic antibiotic treatment significantly reduced the incidence of recurrent episodes of PID compared to placebo. Additionally, participants in the treatment group had a reduced duration of symptoms, improved clinical outcomes, decreased healthcare utilization, and lower rates of pregnancy complications compared to the placebo group. Despite a few adverse effects reported, the benefits of antibiotic treatment in terms of reduced PID recurrence outweighed the associated risks. In conclusion, the findings of this study support the use of prophylactic antibiotic therapy in high-risk women with PID in order to prevent recurrence and improve overall clinical outcomes. Further research is needed to assess the long-term effects of this approach and to explore alternative treatment options for PID.
A high-resolution analysis of the skin in various dermatological conditions remains essential for accurate diagnoses and improved therapeutic strategies. In this study, we present a novel non-invasive approach utilizing advanced imaging techniques to characterize tissue microarchitecture in dermatopathological specimens. By employing multiphoton microscopy coupled with second harmonic generation and fluorescence lifetime imaging microscopy, we successfully visualized collagen organization, cellular morphology, and metabolic changes within the dermis. Our findings reveal distinctive architectural patterns associated with different skin disorders, facilitating the identification and differentiation of various dermatopathologies. This methodology holds great promise for enhancing the precision and efficiency of dermatological diagnosis in clinical practice.
This study examines the efficacy of utilizing probiotics as adjuvant therapy in pediatric patients with acute gastroenteritis (AGE) to reduce symptom duration and severity. A systematic review and meta-analysis of randomized controlled trials (RCTs) from various databases were conducted. Thirteen RCTs involving 1626 pediatric patients were included for analysis. The findings reveal that probiotics administration significantly decreases the duration of diarrhea in pediatric patients by a mean difference of 17.36 hours (95% confidence interval [CI] -23.10 to -11.63, $P < 0.001$). Furthermore, probiotics demonstrate a notable reduction in the risk of diarrhea lasting beyond 48 hours (relative risk [RR] 0.62, 95% CI 0.51-0.76, $P < 0.001$). Moreover, the severity of diarrhea during the first three days of treatment is significantly lower in the probiotics group compared to controls. No severe adverse events related to probiotics were reported. Therefore, probiotics serve as a promising adjunctive option in managing AGE in pediatric patients to expedite recovery and decrease symptom severity.

Table 13: Some examples of synthetic samples generated using the PubMed dataset