CONTRASTIVE PRIVATE DATA SYNTHESIS VIA WEIGHTED MULTI-PLM FUSION

Tianyuan Zou[1], **Yang Liu**[2,3*], **Peng Li**[1], **Yufei Xiong**[4], **Jianqing Zhang**[5], **Jingjing Liu**[1], **Xiaozhou Ye**[6],**Ye Ouyang**[6], **Ya-Qin Zhang**[1]
[1]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[2]the Hong Kong Polytechnic University, HongKong, China
[3]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[4]Harbin Institute of Technology, Weihai, Shandong, China
[5]Shanghai Jiao Tong University, Shanghai, China [6]AsiaInfo Technologies, Shanghai, China

ABSTRACT

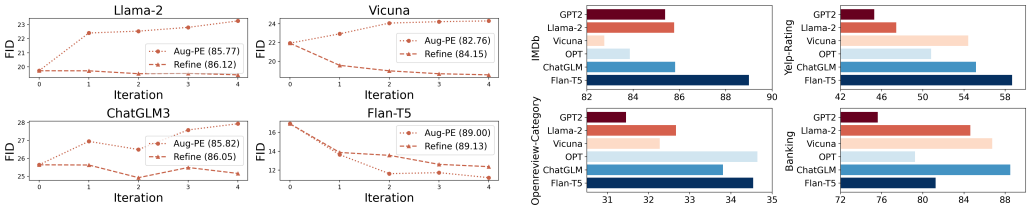
Substantial quantity and high quality are the golden rules of making a good training dataset with sample privacy protection equally important. Generating synthetic samples that resemble high-quality private data while ensuring Differential Privacy (DP), a formal privacy guarantee, promises scalability and practicality. However, existing methods relying on pre-trained models for data synthesis often struggle in data-deficient scenarios, suffering from limited sample size, inevitable generation noise and existing pre-trained model bias. To address these challenges, we propose a novel contr**A**stive private data **S**ynthesis via **W**eighted multiple **P**re-trained language models (PLM) framework, named as **WASP**. WASP utilizes limited private samples for more accurate private data distribution estimation via a Top-Q voting mechanism, and leverages low-quality synthetic samples for contrastive generation via collaboration among dynamically weighted multiple pre-trained models. Extensive experiments on 6 well-developed datasets with 6 open-source and 3 closed-source PLMs demonstrate the superiority of WASP in improving model performance over diverse downstream tasks. Code is available at `https://github.com/LindaLydia/WASP`.

1 INTRODUCTION

In the rapidly evolving landscape of AI models and AI agents, the strength of both Large Language Models (LLMs) and Small Task-specific Models (STMs) hinges on the abundance of high-quality training data Budach et al. (2022); Wang et al. (2024), of which only a limited amount of samples can be harnessed in practice. To further complicate the issue, broad tasks across disciplines such as medical record summarization Rumshisky et al. (2016), chatbots for personalized weight loss Chew (2022) and instruction-following LLM fine-tuning Yu et al. (2024) all rely on high-quality private data collected from real users, which inevitably incurs non-negligible privacy issues.

Differentially private synthetic data stands in as a promising remedy Bommasani et al. (2019); Putta et al. (2023); Flemings & Annavaram (2024), by creating a new synthetic dataset that resembles the real private dataset while preserving the privacy of each sample via guaranteeing Differential Privacy (DP) Dwork (2006). There are two main lines of research for generating DP synthetic datasets. The first line of works Mattern et al. (2022); Yue et al. (2023) introduce DP Fine-tune Generator which involves fine-tuning a Pre-trained Language Model (PLM) using DP-SGD Abadi et al. (2016). However, this practice is computationally intensive and requires substantial data for effective fine-tuning. Another line of work, Private Evolution (PE) Lin et al. (2024); Xie et al. (2024); Hou et al. (2024), relieves the fine-tuning requirement and instead merely uses the APIs of pre-trained models for generation, under DP-protected guidance from private samples. This API-based nature is efficient in creating DP synthetic data, and can leverage both open-source and closed-source pre-trained models, making PE a more applicable solution compared to its counterparts.

*Correspondence to: Yang Liu <yang-veronica.liu@polyu.edu.hk>



(a) (b)

Figure 1: (a) Comparison of the similarity of synthetic dataset to real private dataset (measured by FID Heusel et al. (2017)) and STM performance (numbers within parenthesis) of Aug-PE Xie et al. (2024) (dotted lines) and our refinement with Top-Q voting (dashed lines) under the same DP setting as in Table 1 with IMDb dataset. Lower FID indicates higher similarity. (b) Results of Aug-PE using 100 private samples and under the same DP setting as in Table 1.

Although proven effective, current PE methods Lin et al. (2024); Xie et al. (2024); Hou et al. (2024), still face three major challenges: (1) **Limited Private Samples.** Existing PE methods rely on at least thousands of private samples Lin et al. (2024); Xie et al. (2024); Hou et al. (2024) to guarantee reliable generation feedback selection. In practice, however, data sources may provide only a few hundred samples Zdrazil et al. (2024); Ren et al. (2019), leading to noisy selection guidance. As shown in Figure 1(a), with limited private samples (100), Aug-PE Xie et al. (2024) (PE for text) failed to generate synthetic samples resembling real samples' distribution for 3 PLMs (except Flan-T5). Similar conclusion is drawn in Lin et al. (2024) (see Table 2 therein). This calls for a more precise guidance from limited private samples. (2) **Noisy Synthetic Data.** Although PE approaches encourage the generation of high-quality samples that are close to real private sample distribution, low-quality noisy samples are still unavoidable (see examples in Table 8 in Appendix C.1), which hinder the final performance when training downstream models Ye et al. (2022); Gao et al. (2023); Zou et al. (2024). This highlights the the importance of instructing the avoidance of generating noisy samples during data synthesis. (3) **Risky PLM Selection.** As shown in Figure 1(a), different PLMs yield varying performances (some with unsatisfactory results), and even the best performing model differs across various downstream tasks (see Figure 1(b)), making it hard to select the most suitable pre-trained model for a specific task. Previous PE works primarily focus on single PLM setting, thus the potential of collaboration among multiple PLMs is still unexplored.

To address these demanding challenges, we propose WASP, a collaborative framework that fuses the knowledge from weighted multiple PLMs to synthesize DP data in a contrastive fashion. (1) *To overcome private sample scarcity*, we first extend the voting mechanism for private distribution estimation used in PE from Top-1 to Top-Q ($Q > 1$) with decaying weights, in order to get a more accurate estimation while guaranteeing private data DP. (2) *To reduce noise*, we further leverage the previous voting results to select both high-quality and low-quality samples, and incorporate a contrastive prompt containing both types of samples to improve synthetic data quality by encouraging generation that is more aligned to high-quality samples and less similar to low-quality ones. Notice that under multi-PLM setting, these samples may originate from different PLMs. (3) *To mitigate model bias*, we then interfuse the capabilities of multiple PLMs with dynamically adjusted importance weight for each PLM based on the ensemble votes from private samples. The underlying principle is to assign higher weights to PLMs that generate synthetic samples that are closer to real samples on average. Operating in an iterative fashion, the WASP framework can generate large quantity of synthetic data that better approximate real private data distribution while observing differential privacy. Notably, this process incurs no additional API queries compared to its single-PLM counterparts.

Our contributions are summarized as follows:

1) We introduce a privacy-preserving collaborative framework WASP to facilitate collaboration between multiple PLMs and private samples, especially benefiting scenarios with limited private data.

2) Our proposed WASP leverages differentially private Top-Q voting to improve the estimation of private distributions using limited private samples. It generates higher-quality data by contrasting high- and low-quality samples and dynamically assigns importance weights to PLMs, ensuring that more capable PLMs of the specific task are prioritized.

3) Experiments on 6 well-defined natural language processing tasks with 6 open-source and 3 closed-source PLMs demonstrate the consistent superiority of WASP over existing methods, especially for challenging tasks.

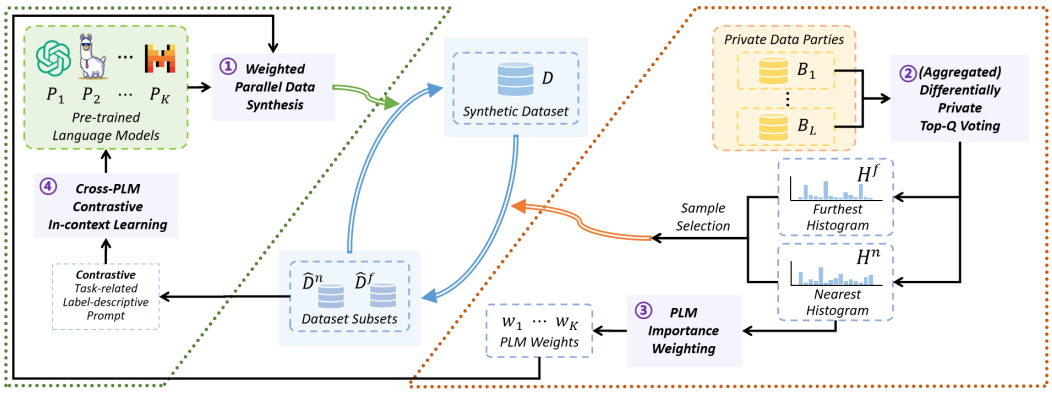


Figure 2: Overview of WASP framework.

2 Related Work

DP Synthetic Dataset. The goal of generating DP synthetic data is to mimic private dataset while protecting sensitive information. Although fine-tuning a PLM with DP-SGD Abadi et al. (2016) for data generation purpose can be effective Bommasani et al. (2019); Putta et al. (2023); Flemings & Annavaram (2024); Mattern et al. (2022); Yue et al. (2023), it is computationally intensive and requires a large number of high-quality private samples to reach strong performance. Moreover, many state-of-the-art PLMs such as GPT series OpenAI (2021; 2023); Hurst et al. (2024) are also closed-source, making DP fine-tuning impractical.

A new line of work instead relies on generative APIs of PLMs without fine-tuning, which focuses on either iterative data synthesis under DP guidance Lin et al. (2024); Zhao et al. (2024); Bojkovic & Loh (2024) or creating DP replica of a given large private dataset Nagesh et al. (2024). Given that requiring a large global dataset for synthetic data initialization Zhao et al. (2024) is hard to obtain in most cases, Lin et al. (2024) proposes a more practical solution, Private Evolution (PE), which instead uses task-related synthetic samples. In PE, private samples are used to identify their nearest synthetic counterparts under DP protection, which then guide the growth of the DP synthetic dataset. PE is proven effective across images Lin et al. (2024) and text Xie et al. (2024), and is further adapted to federated private data scenarios Hou et al. (2024). However, all these works primarily focus on using a single PLM as the generation model.

PLM Fusion. The combination of multiple PLMs can lead to stronger model performance Liu et al. (2024); Du et al. (2023); Wan et al. (2024a;b); Li et al. (2024); Zou et al. (2024). Some studies fine-tune target models with token-level fusion from PLMs as teachers during training time Wan et al. (2024a;b), while others use majority voting Li et al. (2024) or logits averaging Mavromatis et al. (2024) for knowledge fusion during inference. However, data privacy challenges persist, as training or test samples are exposed to external models. To solve this, FuseGen Zou et al. (2024) recently proposes PLM fusion in a zero-shot learning setting, utilizing only model APIs to synthesize data without accessing real private samples, thereby ensuring data privacy. However, by treating all PLMs equally, it overlooks the capability difference of individual PLMs over different tasks.

More related works considering Contrastive In-context Learning are included in Appendix F.

3 Preliminaries

Differential Privacy (DP). If two datasets $\mathcal{D}$ and $\mathcal{D}'$ differ in a single entry, they are referred to as *Neighboring Datasets*. A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any neighboring datasets $\mathcal{D}, \mathcal{D}'$ and any output subset E of $\mathcal{M}$, it holds that Dwork (2006):

$$\Pr[\mathcal{M}(\mathcal{D}) \in E] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in E] + \delta. \tag{1}$$

Note that post-processing on the output of (ϵ, δ)-DP does not incur additional privacy loss Dwork et al. (2014).

Gaussian Mechanism. *Gaussian Mechanism* Dwork (2006) can be applied to guarantee (ϵ, δ)-DP, for any $\epsilon > 0, \delta \in (0, 1)$, by adding Gaussian noise following $\mathcal{N}(0, \sigma^2)$ to the transmitted statistics with $\sigma = \Delta \frac{\sqrt{2\log(1.25/\delta)}}{\epsilon}$ and Δ being the sensitivity of $\mathcal{M}$ Balle & Wang (2018).

4 METHODOLOGY

4.1 PROBLEM DEFINITION

In this paper, we aim to generate a DP synthetic dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of size N using a small number of private data $\mathcal{B} = \{(\mathbf{z}_j, u_j)\}_{j=1}^M$, where M denotes the number of private samples, and $\mathbf{z}_j, u_j$ denote the feature and label of the private sample j, respectively. We consider data-scarcity setting where M is typically a few hundreds at most. To achieve this, we harness the collaborative power of K black-box PLMs $\{\mathcal{P}_k\}_{k=1}^K$ via APIs, while protecting private data by Gaussian DP. For evaluation, we use $\mathcal{D}$ to train a Small Task-specific Model (STM) m and evaluate model performance on a test dataset $\mathcal{A}$ containing real samples that is never used during training.

Note that our framework can be easily extended to the scenario of distributed federated data where each data source possesses an insufficient amount of private data and collaborates on private tasks with secure aggregation Hou et al. (2024). We present the related details in Section 4.7.

4.2 OVERALL WORKFLOW OF WASP

The overall workflow of WASP is depicted in Figure 2 and Algorithm 1, where four steps are taken iteratively for T times. For a given task, the first iteration begins by prompting each PLM $\mathcal{P}_k$ with a zero-shot prompt, which describes the task and category label, to generate a synthetic data subset $\mathcal{D}_k$ of equal size N_k. These samples do not contain information about $\mathcal{B}$. The collective dataset $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$ is then voted by each private sample using a differentially private Top-Q voting mechanism to identify high-quality and low-quality synthetic samples based on their similarity to the distribution of $\mathcal{B}$. These samples are then used to create a contrastive in-context learning prompt for the next round of PLM generation. The voting results are further exploit to dynamically adjust the importance weight w_k for each PLM $\mathcal{P}_k$, which determines N_k of the next generation round. The process repeats from here, expanding $\mathcal{D}$ with DP synthetic samples. After T iterations, $\mathcal{D}$ is used to train an STM m. For notational simplicity, we omit the iteration index t, with $\mathcal{D}$ accumulated over iterations. DP guarantee of WASP is given in Theorem 4.1 with proof included in Appendix D.

Theorem 4.1. *WASP (Algorithm 1) guarantees DP with DP budget ϵ.*

4.3 WEIGHTED PARALLEL DATA SYNTHESIS

In this stage (lines 4-6 in Algorithm 1), each PLM $\mathcal{P}_k$ generates $N_k = [(N/T) \times w_k]$ synthetic samples following:

$$\mathbf{x}_i \sim \mathcal{P}_k\left(\cdot|\mathcal{T}(y_i)\right) , \tag{2}$$

where $\{w_k\}_{k=1}^K$ are the weights for $\{\mathcal{P}_k\}_{k=1}^K$, $[\cdot]$ is the rounding function, N is the expected total number of synthetic samples to be generated, and $\mathcal{T}(\cdot)$ is the generation prompt. In the initial iteration, $\mathcal{T}(\cdot)$ is a zero-shot prompt that describes the task and provides category description, with all PLMs receiving equal weights, i.e. $\{w_k = \frac{1}{K}\}_{k=1}^K$. For later iterations, $\mathcal{T}(\cdot)$ is extended to a few-shot contrastive prompt (see Section 4.6) with in-context samples selected in Section 4.4, and $\{w_k\}_{k=1}^K$ dynamically assigned based on each PLM's capability of the specific task (see Section 4.5). The collective synthetic dataset $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$ is then sent to the private data party.

4.4 DIFFERENTIALLY PRIVATE TOP-Q VOTING

As shown in Figure 1(a), with limited real private samples, noisy estimations of the real private data distribution cause the original PE algorithm to fail in generating synthetic samples that resemble private data. Our aim is to improve distribution estimations and generation guidance in this scenario. To achieve this, unlike previous works Lin et al. (2024); Xie et al. (2024); Hou et al. (2024) that assign only 1 vote per private sample, we propose a Top-Q voting mechanism with decaying weights.

Algorithm 1 WASP

Input:
K PLMs $\{\mathcal{P}_k\}_{k=1}^K$ with empty synthetic dataset $\{\mathcal{D}_k \leftarrow \emptyset\}_{k=1}^K$; 1 data party with private dataset $\mathcal{B}$ of size M belonging to C categories; number of in-context samples S; number of iterations T taken to obtain in total N synthetic samples; initialized PLM weights $\{w_k = 1/K\}_{k=1}^K$; learning rate η; DP privacy parameters $\epsilon, \delta, \delta_{iter}$; test dataset $\mathcal{A}$; random initialized STM $m_{(0)}$.
Output: STM m.

Initialize in-context feedback samples $\hat{\mathcal{D}}^n \leftarrow \emptyset, \hat{\mathcal{D}}^f \leftarrow \emptyset$.
Calculate Gaussian noise $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon}$.
for $t = 0$ **to** $T-1$ **do**
  for $k = 1$ **to** K **in parallel do**
    $\mathcal{D}_k \leftarrow$ `WeightedSynDataGeneration`$(\mathcal{D}_k, \hat{\mathcal{D}}^n, \hat{\mathcal{D}}^f, [(N/T) \times w_k], C)$.
  end for
  $\mathcal{D} \leftarrow \cup_{k=1}^K \mathcal{D}_k$.
  $H^n, H^f \leftarrow$ `DP_PrivateVoting`$(\mathcal{D}, \mathcal{B}, Q, \sigma)$.
  $\hat{\mathcal{D}}^n, \hat{\mathcal{D}}^f \leftarrow$ `SampleSelection`$(\mathcal{D}, H^n, H^f, S, C)$.
  $\{w_k\}_{k=1}^K \leftarrow$ `PLMScoring`$(H^n, \{\mathcal{D}_k\}_{k=1}^K)$.
end for
$m \leftarrow$ `STMTraining`$(\mathcal{D}, m_{(0)}, \eta)$.

This approach maximizes the use of limited private samples by giving weighted votes to the Top-Q nearest and furthest synthetic samples relative to the private sample. Specially, we first compute the pair-wise distance between each of the private samples $(\mathbf{z}_j, u_j) \in \mathcal{B}$ and each synthetic sample $(\mathbf{x}_i, y_i) \in \mathcal{D}$ if they possess the same label, i.e. $y_i = u_j$. Using ℓ_2 distance as measurement, we have:

$$d(\mathbf{z}_j, \mathbf{x}_i) = ||\varphi(\mathbf{z}_j) - \varphi(\mathbf{x}_i)||_2\,,\ \forall\ j = 1, \ldots, M;\ (\mathbf{x}_i, y_i) \in \mathcal{D}^{[u_j]}\,, \tag{3}$$

where φ denotes a pre-trained sentence embedding model and $\mathcal{D}^{[u_j]}$ denotes the subset of $\mathcal{D}$ which has a label that equals to u_j. Next, we use each private sample $(\mathbf{z}_j, u_j) \in \mathcal{B}$ to vote for its Top-Q nearest and Top-Q furthest synthetic samples within $\mathcal{D}^{[u_j]}$ based on Equation (3). The indices of the synthetic samples selected by each $(\mathbf{z}_j, u_j) \in \mathcal{B}$ are:

$$\begin{aligned}
[n_{j,1}, \ldots, n_{j,Q}] &\leftarrow \arg\operatorname{top}Q\mathrm{Smallest}\left(d(\mathbf{z}_j, \mathbf{x}_i)_{(\mathbf{x}_i, y_i) \in \mathcal{D}^{[u_j]}}\right), \\
[f_{j,1}, \ldots, f_{j,Q}] &\leftarrow \arg\operatorname{top}Q\mathrm{Largest}\left(d(\mathbf{z}_j, \mathbf{x}_i)_{(\mathbf{x}_i, y_i) \in \mathcal{D}^{[u_j]}}\right).
\end{aligned} \tag{4}$$

where functions $\arg\operatorname{top}Q\mathrm{Smallest}$ and $\arg\operatorname{top}Q\mathrm{Largest}$ return the indices of the Top-Q samples with the smallest and largest $d(\mathbf{z}_j, \mathbf{x}_i)$, respectively, with $n_{j,1}, \ldots, n_{j,Q}, f_{j,1}, \ldots, f_{j,Q}$ denoting the index of selected samples. To utilize the relative ranking information, as well as to guarantee a controllable function sensitivity for DP protection, we assign decreasing voting weights $1, \frac{1}{2}, \ldots, \frac{1}{2^{Q-1}}$ to each of the Top-Q selected samples when producing the voting histograms, *Nearest Histogram* H^n and *Furthest Histogram* H^f. This can be formulated as:

$$\begin{gathered}
H^n[n_{j,q}] \leftarrow H^n[n_{j,q}] + \frac{1}{2^{q-1}},\ H^f[f_{j,q}] \leftarrow H^f[f_{j,q}] + \frac{1}{2^{q-1}} \\
\forall (\mathbf{z}_j, u_j) \in \mathcal{B},\ \forall\, q \in [1, \ldots, Q],
\end{gathered} \tag{5}$$

with H^n, H^f each initialized as $[0, \ldots, 0]$ of length $|\mathcal{D}|$.

To further guarantee (ϵ, δ)-DP for private samples, Gaussian noises following $\mathcal{N}(0, \sigma^2)$ with $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon}$ are added to H^n, H^f:

$$H^n \leftarrow H^n + \mathcal{N}(0, \sigma^2 I_{|\mathcal{D}|}),\ H^f \leftarrow H^f + \mathcal{N}(0, \sigma^2 I_{|\mathcal{D}|})\,, \tag{6}$$

where $I_{|\mathcal{D}|}$ represents the identity matrix of size $|\mathcal{D}| \times |\mathcal{D}|$ and $\delta_{iter} < \frac{\delta}{(T-1)}$ represents the DP hyper-parameter applied within each iteration.

Based on H^n, H^f, for each category c, we select low-quality samples with the highest votes in H^f (largest distance to private samples in $\mathcal{B}$), denoted as $\hat{\mathcal{D}}^{f,[c]}$, alongside high-quality samples with the

highest votes in H^n (nearest to private samples in $\mathcal{B}$), denoted as $\hat{\mathcal{D}}^{n,[c]}$, following:

$$
\begin{aligned}
H^{n,[c]} &= \left\{ H^n[i] \,\middle|\, (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \right\}, H^{f,[c]} = \left\{ H^f[i] \,\middle|\, (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \right\}, \\
\hat{\mathcal{D}}^{n,[c]} &= \left\{ (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \,\middle|\, H^n[i] \text{ is among the top-}S \text{ values of } H^{n,[c]} \right\}, \\
\hat{\mathcal{D}}^{f,[c]} &= \left\{ (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \,\middle|\, H^f[i] \text{ is among the top-}S \text{ values of } H^{f,[c]} \right\},
\end{aligned} \tag{7}
$$

where S is the amount of samples to select and $H^{n,[c]}, H^{f,[c]}$ denote the sets of the nearest and furthest voting results of samples belonging to category c. $\hat{\mathcal{D}}^n = \bigcup_{c=1}^{C} \hat{\mathcal{D}}^{n,[c]}, \hat{\mathcal{D}}^f = \bigcup_{c=1}^{C} \hat{\mathcal{D}}^{f,[c]}$ are the total sets of high- and low-quality samples respectively. Note that we do not limit the origin of the selected samples, and synthetic samples generated by different PLMs can all be included in $\hat{\mathcal{D}}^n$ and $\hat{\mathcal{D}}^f$.

4.5 PLM Importance Weighting

Previous studies on API-based multi-PLM fusion Li et al. (2024); Zou et al. (2024) often treat involved PLMs equally. However, as shown in Figure 1(b) and Figure 6 in Appendix E.2, different PLMs exhibit varying generation capabilities, leading to uneven synthetic data quality. This encourages assigning customized weights for each PLM to enhance their contributions. Therefore, we introduce a PLM weighting strategy based on the quality of their generated synthetic data, which is measured by their similarity to private samples.

Since the *Nearest Histogram* H^n obtained in Equation (5) quantifies the similarity between each synthetic sample and private samples, we simply aggregate the histogram values of each synthetic sample with source PLM $\mathcal{P}_k$ to obtain the weight w_k of the PLM $\mathcal{P}_k$ for the upcoming generation iteration. That is,

$$
s_i = \frac{H^n[i]}{\sum_{i'=1}^{|\mathcal{D}|} H^n[i']}, w_k = \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} s_i}{\frac{|\mathcal{D}_k|}{\sum_{k'=1}^{K} |\mathcal{D}_{k'}|}} = \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} s_i}{|\mathcal{D}_k|/|\mathcal{D}|}. \tag{8}
$$

4.6 Cross-PLM Contrastive In-context Learning (ICL)

Inspired by the observation that low-quality samples still exist in DP synthetic dataset given by PE (see Table 8 in Appendix C.1), we select cross-PLM contrastive samples from $\hat{\mathcal{D}}^n$ and $\hat{\mathcal{D}}^f$ (obtained in Section 4.5), and use them to create a *contrastive task-related label-descriptive prompt* $\mathcal{T}(\cdot)$ to perform cross-PLM contrastive ICL. $\mathcal{T}(\cdot)$ describes the task, provides category description, and contains explicit contrastive instructions for high- and low-quality samples. It contains the following sequential instructions: (1) analyze the difference between low- and high-quality samples; (2) ensure the new sample is better in quality and closer to real private distribution than the high-quality samples, and is further away from the low-quality samples than the high-quality samples; (3) generate a new sample which is diverse in expression compared to the given high-quality samples. Note that to improve the generation diversity, for each generation we perform random sample selection to draw 50% of samples respectively from $\hat{\mathcal{D}}^{f,[c]}$ and $\hat{\mathcal{D}}^{n,[c]}$ to construct the final in-context samples for $\mathcal{T}(c)$. Also, different from PE algorithms series, we choose not to vary one existed synthetic sample each time, but to encourage diverse sample generation using S demonstrations at once. Prompt examples can be found in Table 7 in Appendix A.

4.7 WASP in Federated Data Setting

So far we have built our algorithms under single data-party setting, which can be easily extended to federated data scenario Hou et al. (2024), where each data party possesses an insufficient amount of private data and collaborates on private tasks. This scenario is very common in the real world, such as collaborations between medical companies. In this setting, we consider L data parties $\{\mathcal{C}_l\}_{l=1}^{L}$, each possessing a real private dataset $\mathcal{B}_l = \{(\mathbf{z}_{l,j}, y_{l,j})\}_{j=1}^{M_l}$ of size M_l. These data parties aim to collaboratively generate a DP synthetic dataset while preserving local data privacy. The full algorithm is provided in Algorithm 2.

When extending to federated data setting, each party $\mathcal{C}_l$ uses its local private samples in $\mathcal{B}_l$ to perform DP Top-Q voting with $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon\sqrt{L}}$ to guarantee privacy. The produced local nearest and furthest voting histograms $\{H_l^n\}_{l=1}^L, \{H_l^f\}_{l=1}^L$ are then securely aggregated Bonawitz et al. (2016) before sent to a central server following:

$$H^n \leftarrow \sum_{l=1}^{L} H_l^n,\ H^f \leftarrow \sum_{l=1}^{L} H_l^f. \tag{9}$$

We adopt an honest-but-curious threat model where the server only has access to the aggregated histograms H^n and H^f, but not individual ones. We also assume that all data parties participate in the aggregation and therefore aims to ensure sample-level (ϵ, δ)-DP of $\mathcal{B}$ (see Appendix D). Note that, WASP can be easily extended to ensure user-level DP, with discussions and results included in Appendix E.4.

5 Experiments

5.1 Settings

Models. In this work, 6 open-source PLMs and 3 closed-source PLMs are considered. Open-source PLMs include GPT-2-xl (GPT-2) Radford et al. (2019), Llama-2-7b-chat-hf (Llama-2) Touvron et al. (2023), Vicuna-7b-1.5v (Vicuna) Chiang et al. (2023), OPT-6.7b (OPT) Zhang et al. (2022), ChatGLM3-6b-base (ChatGLM3) Du et al. (2022), and Flan-T5-xl (Flan-T5) Chung et al. (2022). Close-source PLMs include GPT-3.5-turbo-instruct (GPT-3.5) OpenAI (2021), GPT-4-turbo-preview (GPT-4) OpenAI (2023), and GPT-4o Hurst et al. (2024). For STM, we use pre-trained bert-base-uncased (BERT) model and further fine-tune it on downstream classification tasks using $\mathcal{D}$. We use sentence-t5-base Ni et al. (2022) as the embedding model φ.

Datasets. We evaluate on 6 widely used tasks: 1) IMDb Maas et al. (2011) (2 categories) for movie-review semantic analysis task; 2) Yelp-Category Inc. Yelp (2015) (10 categories) for business-review item field classification task; 3) Yelp-Rating Inc. Yelp (2015) (5 categories) for business-review rating classification task; 4) Openreview-Category Xie et al. (2024) (12 categories) for paper-review classification by research area task; 5) Openreview-Rating Xie et al. (2024) (5 categories) for paper-review classification by review rating task; and 6) Banking (10 categories selected from Banking77 Casanueva et al. (2020)) for online-banking queries field classification task. $\mathcal{B}$ is randomly drawn from the training sets of these datasets with their test sets used to evaluate trained STM.

Baselines. We compare the WASP framework to 4 baselines: 1) Aug-PE Xie et al. (2024), the original PE algorithm specialized for text modality; 2) Pre-Text Hou et al. (2024), which applies PE to federated private data setting; 3) OnlyPrivate, the centralized training method relying merely on $\mathcal{B}$ without DP ($\epsilon = \infty$), which provides a performance upper-bound of using no synthetic data; 4) FuseGen Zou et al. (2024), which generates synthetic data in a zero-shot manner without accessing private samples. "Absolutely Private" in result tables indicates that no private sample in exploit during training.

Implementation Details. By default, we use 100 private samples ($M = 100$) for main experiments. For federated data ($L > 1$) scenario, we use $L = 10$ private data parties which control 300 private samples ($M = \sum_{l=1}^{10} |\mathcal{B}_l| = 300$) altogether. To better align with real-world scenarios, each participating data-party controls private datasets that are *non-i.i.d.* to each other, and aggregate to an unbalanced dataset. We follow Dirichlet Partition Yurochkin et al. (2019); Hsu et al. (2019); Zhang et al. (2023) to distribute private samples to each party with parameter $\alpha = 1.0$. For the DP synthetic dataset, we generate a total of $6,000$ samples from all participating PLMs within 5 iteration. Since the first iteration does not use private sample information for feedback, only the last 4 iterations are sensitive to privacy. By default, we use $\delta_{iter} = 1 \times 10^{-5}$ in our experiments and list only ϵ alongside the results. The notion of DP is sample-level DP unless otherwise stated.

5.2 Main Results

Single Data Party Setting. Experimental results using $K = 6$ open-source PLMs and 3 closed-source PLMs are provided in Tables 1 and 3, which show that WASP outperforms all baseline methods

Table 1: Evaluation of downstream STM accuracy using 6 PLMs, $L = 1$. **Best** and second best results are marked.

		Privacy	$\lvert\mathcal{B}\rvert$	$\lvert\mathcal{D}\rvert$	IMDb	Yelp		Openreview		Banking
						Category	Rating	Area	Rating	
OnlyPrivate		$\epsilon = \infty$	100	-	50.00	5.69	35.57	6.56	22.20	13.75
FuseGen		Absolutely Private	-	$6,000$	89.07	63.38	57.96	24.70	34.57	78.75
Aug-PE	GPT-2	$\epsilon = 4.0$	100	$6,000$	85.38	62.33	45.28	31.45	24.12	75.63
	Llama-2	$\epsilon = 4.0$	100	$6,000$	85.77	60.18	47.42	32.67	34.78	84.63
	Vicuna	$\epsilon = 4.0$	100	$6,000$	82.76	63.28	54.42	32.27	30.66	86.75
	OPT	$\epsilon = 4.0$	100	$6,000$	83.86	62.71	50.81	34.64	25.30	79.25
	ChatGLM3	$\epsilon = 4.0$	100	$6,000$	85.82	55.06	55.17	33.81	32.49	88.50
	Flan-T5	$\epsilon = 4.0$	100	$6,000$	89.00	62.06	58.69	34.54	35.42	81.25
WASP (Ours)		$\epsilon = 4.0$	100	$6,000$	**89.52**	**63.91**	**61.21**	**34.99**	**37.10**	**88.75**

Table 2: Evaluation of downstream STM accuracy using 6 PLMs, $L = 10$. **Best** and second best results are marked.

		Privacy	$\lvert\mathcal{B}\rvert$	$\lvert\mathcal{D}\rvert$	IMDb	Yelp		Openreview		Banking
						Category	Rating	Area	Rating	
OnlyPrivate		$\epsilon = \infty$	100	-	50.00	5.90	38.76	8.86	23.55	16.75
FuseGen		Absolutely Private	-	$6,000$	89.07	63.38	57.96	24.70	34.57	78.75
Pre-Text	GPT-2	$\epsilon = 4.0$	100	$6,000$	85.87	62.58	46.25	37.13	24.45	76.25
	Llama-2	$\epsilon = 4.0$	100	$6,000$	86.09	60.20	51.11	34.24	36.24	85.38
	Vicuna	$\epsilon = 4.0$	100	$6,000$	83.52	64.11	54.76	36.38	30.88	86.13
	OPT	$\epsilon = 4.0$	100	$6,000$	83.98	63.65	52.44	37.67	24.73	79.75
	ChatGLM3	$\epsilon = 4.0$	100	$6,000$	86.32	60.24	56.94	38.14	33.35	89.38
	Flan-T5	$\epsilon = 4.0$	100	$6,000$	89.02	62.82	61.04	38.31	36.53	81.75
WASP (Ours)		$\epsilon = 4.0$	100	$6,000$	**89.65**	**64.34**	**61.46**	**40.47**	**37.60**	**89.63**

across different tasks, demonstrating its superiority. As expected, the closed-source GPT series (see Table 3), being powerful models, outperform their open-source counterparts (see Table 1) when using baseline method Aug-PE.

For all tasks, with limited private samples, Aug-PE performs poorly when using improper single PLM, e.g. using OPT for IMDb and using GPT-2 for Openreview-Rating. Differently, WASP performs consistently well across tasks, and achieves a lower FID value compared to baselines (see Figure 5 in Appendix E.1), verifying its effectiveness under limited private sample setting. Also, the best performing PLM model varies across tasks for Aug-PE, highlighting the arbitrary nature of PLM selection. In contrast, WASP consistently achieves better performance across tasks, making it PLM-agnostic without requiring prior-knowledge for selecting specific PLMs for collaboration.

On the other hand, comparing with FuseGen, a baseline under zero-shot setting where private samples are inaccessible, WASP leverages real private samples and utilizes a more targeted PLM importance weighting method, therefore achieving better performance. Moreover, the notably poor performance of "OnlyPrivate" shows that the trained STM relying merely on private dataset $\mathcal{B}$ is nearly unusable, even without applying DP during training which can further degrade STM performance.

Federated Data Setting. We also conduct experiments under distributed federated data setting, with $L = 10$ and $M = 300$ total number of private samples. Results in Table 2 show that WASP consistently achieves better performance across different tasks and settings compared to Pre-Text,

Table 3: Evaluation of downstream STM accuracy using 3 closed-source PLMs, $L = 1$ with the same DP setting in Table 1. **Best** and second best results are marked.

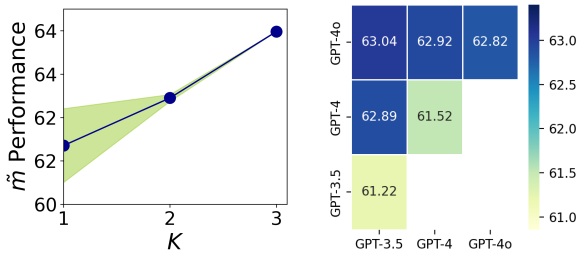
	Only Private	FuseGen	Aug-PE			WASP (Ours)
			GPT-3.5	GPT-4	GPT-4o	
Yelp-Rating	35.57	61.36	60.90	61.02	62.06	**64.48**

a baseline designed for federated data. This further demonstrates the effectiveness of WASP when extended to federated data setting. Additional results on communication cost comparison is given in Table 11 in Appendix E.5, where we show that the communication increase caused by uploading additional histograms by our method is minimal.

5.3 Ablation Studies

 # PLMs (K). We first study the impact of the number of PLMs (K) on the final STM performance. Results of using $1, 2, 3$ closed-source PLMs under $L = 1$ and the same DP setting as in Table 1 are reported in Figure 3(a). We can see that the performance of m increases simultaneously with the increase of K while the randomness (STD) decreases. This indicates that the randomness in the performance of the synthetic dataset can be mitigated by incorporating more PLMs into WASP, which simultaneously increases the performance expectations.

We also display the pair-wise combination ($K = 2$) results of the 3 closed-source PLMs under $L = 1$ and the same DP setting as in Table 1 in Figure 3(b). In this figure, any pair-wise collaboration ($K = 2$) outperforms either participating single-PLM alone (diagnose in Figure 3(b)), demonstrating that WASP performs better using the whole set of available PLMs than using only a subset of them. These findings show that WASP's improvements are PLM-agnostic, independent of any single PLM's inherent task capabilities. Consequently, WASP effectively mitigates the risk of selecting the optimal PLM by harnessing the collective strengths of all participating models.



(a) Effect of K (b) Pair-wise Comparison

Figure 3: Evaluation of downstream STM accuracy using Yelp-Rating dataset with $K = 1, 2, 3$ closed-source PLMs, $L = 1$ under the same DP setting as in Table 1. In (b), results on the diagnose are with $K = 1$ and others are with $K = 2$.

Contrastive ICL & PLM Importance Weighting. To evaluate the effectiveness of our proposed *Contrastive In-context Learning* and *PLM Importance Weighting* methods, we conduct ablation experiments to see how these components impact the final STM performance. Results are reported in Table 4. By removing *Contrastive In-context Learning* (labeled as "w/o PLM Contrastive Prompting" in Table 4), we only select high-quality samples for the prompt (details in Table 7). This leads to a 0.31% decrease in STM performance on the easier IMDb task, and a much larger 1.56% and 0.92% decrease on the more challenging Yelp-Rating and Openreview-Rating tasks. This highlights the importance of using low-quality samples as feedback demonstrations to encourage the PLMs avoid generating low-quality DP synthetic samples.

On the other hand, by removing *PLM Importance Weighting* (labeled as "w/o PLM Importance Weighting" in Table 4), $w_k = 1/K$ within each generation iteration, indicating that each PLM generates equal amount of samples across iterations. Similarly, results indicate a 0.35% decrease in STM performance on the easier IMDb task and a 2.27% and 1.57% decline on the more challenging Yelp-Rating and Openreview-Rating tasks. This underlines the effectiveness of weighted aggregation of PLMs with varying degrees of reliance on their capabilities for specific task. Furthermore, these results demonstrate that by generating better DP synthetic data, WASP is more effective than baselines when faced with more challenging tasks.

 # Votes (Q) by Each Private Sample. To better estimate private sample distribution with limited private samples, WASP exploits each private sample by increasing the amount of votes each private sample gives out (from $Q = 1$ in previous works to $Q = 8$). Here we investigate how the change in Q impacts STM performance. Results in Table 5 indicate that STM performance improves with higher values of Q, but the improvement diminishes at larger Q ($Q > 8$). This underscores the strength of our idea in increasing the utility of each private sample to achieve a more accurate private sample distribution estimation, particularly in scenarios with limited available private samples.

Table 4: Comparison of downstream STM accuracy under w/ and w/o Contrastive In-context Learning and Private Data Assisted PLM Importance Weighting setting using 6 open-source PLMs, $L = 1$ with the same DP setting as in Table 1.

	w/o PLM Contrastive Prompting	w/o PLM Importance Weighting	WASP (Ours)
IMDb	89.21	89.17	89.52
Yelp-Rating	59.65	58.94	61.21
Openreview-Rating	36.18	35.53	37.10

Table 5: Evaluation of downstream STM accuracy using 6 open-source PLMs, $L = 1$ with the same DP setting as in Table 1 under different Q.

	$Q = 1$	$Q = 2$	$Q = 4$	$Q = 8$	$Q = 16$
IMDb	89.02	89.15	89.39	89.52	89.60
Yelp-Rating	58.74	58.92	59.24	61.21	61.42

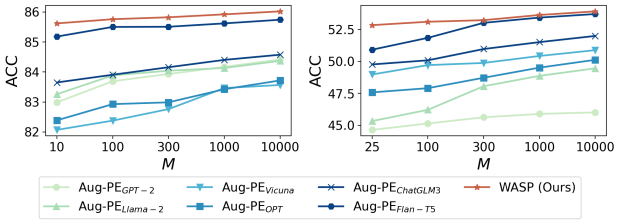


Figure 4: Comparison of downstream STM accuracy using different number of private samples (M) from the training set of IMDb and Yelp-Rating datasets using 6 open-source PLMs, $L = 1$ with the same DP setting as in Table 1.

Sensitive Analysis of # Private Samples (M). We also investigate the impact of M on WASP. In Figure 4, we compare the baseline Aug-PE method with WASP across different values of M. Results show that WASP consistently outperforms Aug-PE across different values of M for all PLMs, and the performance gaps at smaller M values ($M < 1000$) are much greater, underscoring the effectiveness of WASP in limited private data scenarios.

Different Private Budget (ϵ). As illustrated in Table 6, STM performance using WASP gradually declines from 89.96% to 89.36% for IMDb and from 62.02% to 60.94% for Yelp-Rating as the privacy budget ϵ decreases from $\infty, 8.0, 4.0$ to 1.0, similar to that of Aug-PE when using the best performing single PLM for each task. This indicates that WASP scales well with ϵ and maintains high performance even under tight privacy constraints, just like baseline method.

6 Conclusion and Future Work

In this work, we introduce a novel DP synthetic data generation framework, WASP, which leverages the collaborative capabilities of multiple PLMs to address real-world scenarios with limited private samples, while observing differential privacy. Extensive experiments across 6 tasks demonstrate that WASP is highly effective, PLM-agnostic, scalable with respect to privacy budgets, and superior in challenging scenarios, making it a practical and scalable solution for real-world applications.
Possible future work points to more precise sample-level weighting or selection to further improve the quality of the DP synthetic dataset, as well as verifying the effectiveness of WASP on non-classification tasks.

Table 6: Evaluation of downstream STM accuracy using 6 open-source PLMs, $L = 1$ under different DP budget setting with $\delta_{iter} = 1 \times 10^{-5}$. The best performing PLM is used for Aug-PE evaluation, i.e. Flan-T5 for both tasks.

		$\epsilon = \infty$	$\epsilon = 8.0$	$\epsilon = 4.0$	$\epsilon = 1.0$
IMDb	WASP	89.96	89.77	89.52	89.36
	Aug-PE	89.48	89.23	89.00	88.72
Yelp-Rating	WASP	62.02	61.54	61.21	60.94
	Aug-PE	59.62	59.12	58.69	58.59

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant No.2022ZD0160504, the Tsinghua University (AIR)-Asiainfo Technologies (China) Inc. Joint Research Center.

REFERENCES

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

Borja Balle and Yu-Xiang Wang. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Nikolija Bojkovic and Po-Ling Loh. Differentially Private Synthetic Data with Private Density Estimation. *arXiv preprint arXiv:2405.04554*, 2024.

Rishi Bommasani, Steven Wu, and Xanda Schofield. Towards Private Synthetic Text Generation. In *NeurIPS 2019 Machine Learning with Guarantees Workshop*, 2019.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Federated Learning on User-Held Data. In *NIPS 2016 Workshop on Private Multi-Party Machine Learning*, 2016.

Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Hazar Harmouch, and Felix Naumann. The Effects of Data Quality on ML-Model Performance. *CoRR abs/2207.14529*, 2022.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, mar 2020. URL https://arxiv.org/abs/2003.04807. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Han Shi Jocelyn Chew. The Use of Artificial Intelligence–based Conversational Agents (Chatbots) for Weight Loss: Scoping Review and Practical Recommendations. *JMIR medical informatics*, 10 (4):e32578, 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling Instruction-finetuned Language Models. *arXiv preprint arXiv:2210.11416*, 2022.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325*, 2023.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

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

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

James Flemings and Murali Annavaram. Differentially Private Knowledge Distillation via Synthetic Text Generation. *arXiv preprint arXiv:2403.00932*, 2024.

Jiahui Gao, Renjie Pi, Lin Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided Noise-free Data Generation for Efficient Zero-shot Learning. In *Proceedings of The Eleventh International Conference on Learning Representations*, 2023.

Xiang Gao and Kamalika Das. Customizing Language Model Responses with Contrastive In-Context Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18039–18046, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. PrE-Text: Training Language Models on Private Federated Data in the Age of LLMs. In *Forty-first International Conference on Machine Learning*, 2024.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the Effects of Non-identical Data Distribution for Federated Visual Classification. *arXiv preprint arXiv:1909.06335*, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024.

Inc. Yelp. Yelp Open Dataset, 2015. URL `https://www.yelp.com/dataset`.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More Agents is All You Need. *arXiv preprint arXiv:2402.05120*, 2024.

Chao Liang, Wei Xiang, and Bang Wang. In-context Contrastive Learning for Event Causality Identification. *arXiv preprint arXiv:2405.10512*, 2024.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially Private Synthetic Data via Foundation Model APIs 1: Images. In *The Twelfth International Conference on Learning Representations*, 2024.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. In *Proceedings of the 1st Conference on Language Modeling (COLM 2024).*, 2024.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. Differentially Private Language Models for Secure Data Sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4860–4873, 2022.

Costas Mavromatis, Petros Karypis, and George Karypis. Pack of LLMs: Model Fusion at Test-Time via Perplexity Optimization. *arXiv preprint arXiv:2404.11531*, 2024.

Yosuke Miyanishi and Minh Le Nguyen. Multimodal Contrastive In-Context Learning. *arXiv preprint arXiv:2408.12959*, 2024.

Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. C-ICL: Contrastive In-Context Learning for Information Extraction. *arXiv preprint arXiv:2402.11254*, 2024.

Supriya Nagesh, Justin Y Chen, Nina Mishra, and Tal Wagner. Private Text Generation by Seeding Large Language Model Prompts. In *GenAI for Health: Potential, Trust and Policy Compliance*, 2024.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, 2022.

OpenAI. GPT-3.5-Turbo, 2021. URL https://platform.openai.com/docs/models/gpt-3-5-turbo.

OpenAI. GPT-4-Turbo and GPT-4, 2023. URL https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

Pranav Putta, Ander Steele, and Joseph W Ferrara. Differentially Private Conditional Text Generation For Synthetic Data Production, 2023. URL https://openreview.net/forum?id=LUql3ZOFwFD.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

Ruifeng Ren and Yong Liu. Towards Understanding How Transformers Learn In-context Through a Representation Learning Lens. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dB6gwSDXKL.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost Unsupervised Text to Speech and Automatic Speech Recognition. In *International conference on machine learning*, pp. 5410–5419. PMLR, 2019.

Anna Rumshisky, Marzyeh Ghassemi, Tristan Naumann, Peter Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative Discharge Summaries. *Translational psychiatry*, 6(10):e921–e921, 2016.

Thomas Steinke. Composition of Differential Privacy & Privacy Amplification by Subsampling, 2022. URL https://arxiv.org/abs/2210.00597.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-tuned Chat Models, 2023. *URL https://arxiv. org/abs/2307.09288*, 2023.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge Fusion of Large Language Models. In *Proceedings of The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=jiDsk12qcz.

Fanqi Wan, Ziyi Yang, Longguang Zhong, Xiaojun Quan, Xinting Huang, and Wei Bi. FuseChat: Knowledge Fusion of Chat Models. *arXiv preprint arXiv:2402.16107*, 2024b.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. A Survey on Data Selection for LLM Instruction Tuning. *arXiv preprint arXiv:2402.05123*, 2024.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. In *Forty-first International Conference on Machine Learning*, 2024.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3671–3683, 2022.

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-Preserving Instructions for Aligning Large Language Models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=mUT1biz09t.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian Nonparametric Federated Learning of Neural Networks. In *International conference on machine learning*, pp. 7252–7261. PMLR, 2019.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. FedALA: Adaptive Local Aggregation for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11237–11244, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, 2022.

Wenhao Zhao, Shaoyang Song, and Chunlai Zhou. Generate Synthetic Text Approximating the Private Distribution with Differential Privacy. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6651–6659, 2024.

Tianyuan Zou, Yang Liu, Peng Li, Jianqing Zhang, Jingjing Liu, and Ya-Qin Zhang. FuseGen: PLM Fusion for Data-generation based Zero-shot Learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2172–2190, November 2024.

A CONTRASTIVE PROMPTS AND NON-CONTRASTIVE PROMPTS

Table 7: Prompt used for synthetic dataset generation. Due to clarity, we omit the words in the parentheses in the labels of Openreview-Category and the attributes of Openreview-Rating.

Dataset (task)	prompt type	prompt	label	attribute
IMDb (semantic analysis of movie review)	w/o Contrastive	"The movie review is: <*sample_1*>\n\nThe movie review is: <*sample_2*>\n\n...\n\nThe movie review is: <*sample_S*>\n\n\nBased on the above movie reviews, a new movie review also in ***<label>*** sentiment but diverse in the expression compared to the above given samples is: "	***positive / negative***	None
	w/ Contrastive	"A bad movie review is: <*sample_1*>\n\n...\n\nA bad movie review is: <*sample_⌊S/2⌋*>\n\nA good movie review is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good movie review is: <*sample_S*>\n\n\nBased on the above examples of bad and good movie reviews in ***<label>*** sentiment, analyze the differences between the bad and good reviews. Generate a new positive movie review that is diverse in expression compared to the given good reviews. Ensure that the new review is further refined than the good reviews while maintaining the ***<label>*** sentiment and clarity, making the good reviews appear to lie midway between the new review and the bad reviews. The new ***<label>*** movie review is: "		
Yelp-Category (field classification of business review)	w/o Contrastive	The business review is: <*sample_1*>\n\nThe business review is: <*sample_2*>\n\n...\n\nThe business review is: <*sample_S*>\n\n\nBased on the above business reviews belonging to the category of ***<label>***, a new review for a business item also in the field of ***<label>*** with rating <*attribute*> star(s) but diverse in the expression compared to the above given samples is: "	***Arts & Entertainment / Bars / Beauty & Spas / Event Planning & Services / Grocery / Health & Medical / Home & Garden / Hotels & Travel / Restaurants / Shopping***	1.0 / 2.0 / 3.0 / 4.0 / 5.0
	w/ Contrastive	A bad business review is: <*sample_1*>\n\n...\n\nA bad business review is: <*sample_⌊S/2⌋*>\n\nA good business review is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good business review is: <*sample_S*>\n\n\nBased on the above examples of bad and good business reviews belonging to the category of ***<label>***, analyze the differences between the bad and good reviews. Generate a new review for a business item also in the field of ***<label>*** with rating <*attribute*> star(s) but diverse in the expression compared to the given good reviews. Ensure that the new review is further refined than the good reviews while maintaining clarity, making the good reviews appear to lie midway between the new review and the bad reviews. The new business review in the field of ***<label>*** is: "		
Yelp-Rating (rating classification of business review)	w/o Contrastive	The business review is: <*sample_1*>\n\nThe business review is: <*sample_2*>\n\n...\n\nThe business review is: <*sample_S*>\n\n\nBased on the above business reviews with rating ***<label>*** star(s), a new review for a business item in the field of <*attribute*> also with rating ***<label>*** star(s) but diverse in the expression compared to the above given samples is: "	***1.0 / 2.0 / 3.0 / 4.0 / 5.0***	Arts & Entertainment / Bars / Beauty & Spas / Event Planning & Services / Grocery / Health & Medical / Home & Garden / Hotels & Travel / Restaurants / Shopping
	w/ Contrastive	A bad business review is: <*sample_1*>\n\n...\n\nA bad business review is: <*sample_⌊S/2⌋*>\n\nA good business review is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good business review is: <*sample_S*>\n\n\nBased on the above examples of bad and good business reviews with rating ***<label>*** star(s), analyze the differences between the bad and good reviews. Generate a new review for a business item in the field of <*attribute*> also with rating ***<label>*** star(s) but diverse in the expression compared to the above given good reviews. Ensure that the new review is further refined than the good reviews while maintaining clarity, making the good reviews appear to lie midway between the new review and the bad reviews. The new business review with rating ***<label>*** star(s) is: "		
Openreview-Category (field classification of paper review)	w/o Contrastive	The paper review is: <*sample_1*>\n\nThe paper review is: <*sample_2*>\n\n...\n\nThe paper review is: <*sample_S*>\n\n\nBased on the above paper reviews of paper in the area ***<label>***, a new review for a paper also in the area of ***<label>*** with final recommendation: '<*attribute*>' but diverse in the expression compared to the above given samples is: "	***Applications / Deep Learning and representational learning / General Machine Learning / Generative models / Machine Learning for Sciences / Neuroscience and Cognitive Science / Optimization / Probabilistic Methods / Reinforcement Learning / Social Aspects of Machine Learning / Theory / Unsupervised and Self-supervised learning***	1: strong reject / 3: reject, not good enough / 5: marginally below the acceptance threshold / 6: marginally above the acceptance threshold / 8: accept, good paper
	w/ Contrastive	A bad paper review is: <*sample_1*>\n\n...\n\nA bad paper review is: <*sample_⌊S/2⌋*>\n\nA good paper review is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good paper review is: <*sample_S*>\n\n\nBased on the above examples of bad and good paper reviews of paper in the area ***<label>***, analyze the differences between the bad and good reviews. Generate a new review for a paper also in the area of ***<label>*** with final recommendation: '<*attribute*>' but diverse in the expression compared to the given good reviews. Ensure that the new review is further refined than the good reviews while maintaining clarity, making the good reviews appear to lie midway between the new review and the bad reviews. The new paper review in the area ***<label>*** is: "		
Openreview-Rating (rating classification of paper review)	w/o Contrastive	The paper review is: <*sample_1*>\n\nThe paper review is: <*sample_2*>\n\n...\n\nThe paper review is: <*sample_S*>\n\n\nBased on the above paper reviews of final recommendation: ***<label>***, a new review for a paper in the field of '<*attribute*>' also with final recommendation: ***<label>*** but diverse in the expression compared to the above given samples is: "	***1: strong reject / 3: reject, not good enough / 5: marginally below the acceptance threshold / 6: marginally above the acceptance threshold / 8: accept, good paper***	Applications / Deep Learning and representational learning / General Machine Learning / Generative models / Machine Learning for Sciences / Neuroscience and Cognitive Science / Optimization / Probabilistic Methods / Reinforcement Learning / Social Aspects of Machine Learning / Theory / Unsupervised and Self-supervised learning
	w/ Contrastive	A bad paper review is: <*sample_1*>\n\n...\n\nA bad paper review is: <*sample_⌊S/2⌋*>\n\nA good paper review is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good paper review is: <*sample_S*>\n\n\nBased on the above examples of bad and good paper reviews of final recommendation: ***<label>***, analyze the differences between the bad and good reviews. Generate a new review for a paper in the field of '<*attribute*>' also with final recommendation: ***<label>*** but diverse in the expression compared to the above given good reviews. Ensure that the new review is further refined than the good reviews while maintaining clarity, making the good reviews appear to lie midway between the new review and the bad reviews. The new paper review of final recommendation: ***<label>*** is: "		
Banking (field classification of online banking queries)	w/o Contrastive	The online banking query is: <*sample_1*>\n\nThe online banking query is: <*sample_2*>\n\n...\n\nThe online banking query is: <*sample_S*>\n\n\nBased on the above online banking queries in the category of "***<label>***", a new online banking query also in the category of "***<label>***" but diverse in the expression compared to the above given samples is: "	***activate_my_card / age_limit / apple_pay_or_google_pay / atm_support / automatic_top_up / balance_not_updated_after_bank_transfer / balance_not_updated_after_cheque_or_cash_deposit / beneficiary_not_allowed / cancel_transfer / card_about_to_expire***	None
	w/ Contrastive	A bad online banking query is: <*sample_1*>\n\n...\n\nA bad online banking query is: <*sample_⌊S/2⌋*>\n\nA good online banking query is: <*sample_⌊S/2⌋+1*>\n\n...\n\nA good online banking query is: <*sample_S*>\n\n\nBased on the above examples of bad and good online banking queries in the category of "***<label>***", analyze the differences between the bad and good reviews. Generate a new online banking query also in the category of "***<label>***" but diverse in the expression compared to the above given good queries. Ensure that the new query is further refined than the good queries while maintaining clarity, making the good queries appear to lie midway between the new query and the bad queries. The new online banking query also in the category of "***<label>***" is: "		

In Table 7, we listed the prompts used in our experiments, including contrastive ("w Contrastive") and non-contrastive ("w/o Contrastive") in-context learning prompts. We need to clarify that, for PE series baselines, we use their original prompt for `VARIATIONAL_API`, which is different from the

Algorithm 2 WASP for Distributed Federated Data ($L > 1$)

Input:
K PLMs $\{\mathcal{P}_k\}_{k=1}^K$ with empty synthetic dataset $\{\mathcal{D}_k \leftarrow \emptyset\}_{k=1}^K$;
L private data parties controlling distributed private dataset $\{\mathcal{B}_l\}_{l=1}^L$ of M samples in total that belongs to C categories;
number of in-context samples S;
number of iterations T taken to obtain in total N synthetic samples;
initialized PLM weights $\{w_k = 1/K\}_{k=1}^K$;
learning rate η;
DP privacy parameters ϵ, δ;
test dataset of downstream task $\mathcal{A}$;
random initialized STM $m_{(0)}$;
Output: STM m.

1: Initialize in-context feedback samples $\hat{\mathcal{D}}^n \leftarrow \emptyset, \hat{\mathcal{D}}^f \leftarrow \emptyset$.
2: Calculate Gaussian noise $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon\sqrt{L}}$.
3: **for** $t = 0$ **to** $T-1$ **do**
4: **for** $k = 1$ **to** K **in parallel do**
5: $\mathcal{D}_k \leftarrow$ `WeightedSynDataGeneration`$(\mathcal{D}_k, \hat{\mathcal{D}}^n, \hat{\mathcal{D}}^f, [(N/T) \times w_k], C)$.
6: **end for**
7: $\mathcal{D} \leftarrow \cup_{k=1}^K \mathcal{D}_k$.
8: **for** $l = 1$ **to** L **in parallel do**
9: $H_l^n, H_l^f \leftarrow$ `DP_PrivateVoting`$(\mathcal{D}, \mathcal{B}_l, Q, \sigma)$.
10: **end for**
11: $H^n \leftarrow \sum_{l=1}^L H_l^n$; $H^f \leftarrow \sum_{l=1}^L H_l^f$.
12: $\hat{\mathcal{D}}^n, \hat{\mathcal{D}}^f \leftarrow$ `SampleSelection`$(\mathcal{D}, H^n, H^f, S, C)$.
13: $\{w_k\}_{k=1}^K \leftarrow$ `PLMScoring`$(H^n, \{\mathcal{D}_k\}_{k=1}^K)$.
14: **end for**
15: $m \leftarrow$ `STMTraining`$(\mathcal{D}, m_{(0)}, \eta)$.

listed "w/o Contrastive in-context learning" prompt in Table 7. Please refer to Xie et al. (2024) (the original work) for detailed prompts.

B Algorithm for Distributed Private Data and Detailed Functions

Due to space limitation, we include the full algorithm for $L > 1$ setting here in Algorithm 2 in the Appendix. The difference between Algorithm 2 and Algorithm 1 mainly falls in line 2 and lines 8 to 11 in Algorithm 2.

We also include pseudo-code for the functions used in Algorithms 1 and 2 here in Algorithm 3 due to space limitation.

C Supporting Results for Introduction

C.1 Examples of High-quality and Low-quality Samples

We show examples of high-quality and low-quality synthetic samples generated using Aug-PE in Table 9 and Table 8 respectively. We also include the appearance frequency of some types of low-quality samples within the generated DP synthetic dataset in Table 8.

Table 8 shows that, low-quality noisy samples often diverge from the specified task (generating movie reviews in positive/negative sentiment for this table). Differently, likes shown in Table 9, high-quality samples often possess a clear sentiment tendency that well accomplished the task, with some offering detailed judgments or even containing concession details.

Table 8: Low-quality DP synthetic samples for movie review semantic analysis with IMDb as real dataset.

Model	Low-quality Noisy Sample Text (Examples)	Label	Explain
GPT-2	"~ If you are missing No. 17 (see below) ~"	negative	Meaning less sentence.
Llama-2	"In an informal way, please rephrase the sentences as follows:"	positive	Repeating the prompt. Appears around 30 times in $\mathcal{D}$ of size $6,000$.
Vicuna	" \n번역결과 \n좋은 감정을 기반으로 영화 관람 후 즐거움과 엔터테인먼트 가치를 즐기셨기에 좋습니다. "	positive	Unmatched language and translating the prompt. Appears within 9 samples in $\mathcal{D}$ of size $6,000$.
OPT	"The sentence is positive, but is ambiguous. Would it make sense to remove "	negative	Task unrelated and mislabeled.
ChatGLM3	"... 指导意见：句子间不顺畅，需要调整一下顺序，去掉一些冗余的词汇。 Since is based on positive sentiment, then how does one use the 'pronoun?"	positive	Unmatched language and repeating the prompt.
Flan-T5	"what a fantastic movie!! :O great acting and spectacular scenery. great work. i look forward to another good one next year :D come on laura delana!! woohoooo!!!....can't wait!!....."	positive	Excessively long consecutive punctuation marks. Appears within 4 samples in $\mathcal{D}$ of size $6,000$.

Table 9: High-quality DP synthetic samples for movie review semantic analysis with IMDb as real dataset.

Model	High-quality Sample Text (Examples)	Label
GPT-2	"This movie isn't for everyone, but it's worth a look. But if you have to be in the dark to enjoy it, just buy the soundtrack."	positive
GPT-2	"is that it has some semi-interesting world-building, but the information it offers is grossly insufficient to allow one to progress in the story. The central characters have little to do with the world beyond playing around with their toys and talking to themselves. "	negative
Llama-2	"Incredible film! I laughed, I cried, and I left the theater feeling inspired. Every performance was top-notch and the direction was flawless. A true masterpiece that will be remembered for years to come. 10/10."	positive
Llama-2	"This film was a colossal disappointment. I found myself glancing at my wrist every 10 minutes, eagerly anticipating its conclusion. The narrative was intricate and challenging to comprehend, and the personas were underdeveloped. I had no emotional investment in their plight. The performances were lacking, and the dialogue was trite and unconvincing. In summary, I would not advise viewing this motion picture to any individual. "	negative
Vicuna	"A movie for people who want to see an action-packed adventure that will keep them on the edge of their seats."	positive
Vicuna	"Felt like a missed opportunity. Glacial pacing and shallow characters. The premise was intriguing but the execution fell flat. Avoid this one."	negative
OPT	"-This film is enjoyable to watch because of it's well-developed storyline and most of all its storyline that i can relate to and would like to have. This film also has good acting performances and the songs help the film a lot. It's not a typical love story kind of a film, but it does have lots of interesting messages to give. Very well-done! Definitely recommend this film!"	positive
OPT	"Quote: The first sentence means nothing to me. Without context the first two sentences also mean nothing."	negative
ChatGLM3	"Attention getter, visually interesting and outstanding acting, the story of an American citizen that is in Mexico and gets involved in a murder is a good movie. "	positive
ChatGLM3	"This movie is an unwatchable mess, basically a rip off of that scene in Dr. Strangelove where the U.S. Air Force gets into a scramble to deliver nukes after a false alert... I like pretty much every aspect of anime, but this stuff is bad. Animation is choppy, sound effects and voice acting are atrocious, and the plot makes little sense."	negative
Flan-T5	"<pad> This is a fine historical drama about America's Civil Rights movement, that has William Friedkin as director and Gary Cooper as an actor. Although there are other excellent directors that we haven't heard of, William Friedkin has become the standard in many ways. Filming in New Orleans, Mr. Friedkin gives a glimpse into the African-American experience during the late 1920s. I saw it at the All-African Cinema festival at Wharton Business School. I was awestruck by the power and impact of the film. *...[truncated because of excessive length]...*"	positive
Flan-T5	"<pad> beware of this film! i can't recall a single horror movie that was scarier than this one. it's just plain cheesy and stupid. there's no plot and the characters are entirely cheesy. so don't even bother. you'll never forget it."	negative

D THEORETICAL PRIVACY ANALYSIS FOR WASP

To prove Theorem 4.1, in this part, we prove that the WASP framework described in Algorithm 2 with distributed federated data ($L > 1$) satisfies (ϵ, δ)-DP, which is the general case for $L = 1$ setting described in Algorithm 1 and Theorem 4.1.

Theorem D.1. *Let f be a function with global L_2 sensitivity Δ. For any $\epsilon > 0, \delta \in (0, 1)$, the Gaussian output perturbation mechanism with $\sigma = \Delta \frac{\sqrt{2\log(1.25/\delta_{iter})}}{\epsilon}$ ensures that f satisfies (ϵ, δ)-DP.*

Proof of Theorem D.1 can be found in Balle & Wang (2018).

Theorem D.2. *The global L_2 sensitivity Δ of WASP described in Algorithm 2 is 4.*

Proof. In WASP framework (Algorithm 2), function `DP_PrivateVoting` is the only function that accesses the private dataset $\mathcal{B}$. Thus, Δ of WASP equals to that of function `DP_PrivateVoting`. Within function `DP_PrivateVoting`, for nearest histogram and furthest histogram respectively, each private sample contributes Q votes with decaying voting weights $\{1, \frac{1}{2}, \ldots, \frac{1}{2^{Q-1}}\}$. Therefore, the total votes contributed by one private sample is $\sum_{q=1}^{Q} \frac{1}{2^{q-1}} = 2 - \frac{1}{2^{Q-1}} < 2$ for each histogram. Adding or removing one private sample in $\mathcal{B}$ will result in a change no more than 2 in the ℓ_2 norm of each histogram. Therefore, the upper bound of the sensitivity for each histogram is 2 and the upper bound of the sensitivity of WASP framework is 4 considering both histograms, i.e. $\Delta = 4$. □

Lemma D.3. *If a Gaussian mechanism satisfies (ϵ, δ)-DP, then independently repeating this mechanism for T times results in the final DP budget to increase to $\epsilon_{final} = \sqrt{T} \cdot \epsilon$ and the final probability of data leak δ_{final} increased to $\delta_{final} > T \cdot \delta$.*

The proof of Lemma D.3 can be found in Steinke (2022).

With the above lemma and theorems, we present and prove our main theorem as follows.

Theorem D.4. *If each private data party performs standard Gaussian mechanism with addition noise following $\mathcal{N}(0, \sigma^2)$ and $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon\sqrt{L}}$, WASP framework described in Algorithm 2 satisfies DP with privacy budget ϵ for private samples in $\mathcal{B}$.*

Proof. For guaranteeing (ϵ, δ)-DP throughout the $T - 1$ iterations with feedback (the first generation iteration does not use feedback), each iteration should satisfy a differential privacy budget of $\frac{\epsilon}{\sqrt{T-1}}$. Given $\Delta = 4$ for WASP, $\sigma_{total} = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon}$ for each single generation iteration will guarantee (ϵ, δ)-DP for the whole process with $\delta > \delta_{iter} \cdot (T - 1)$. Further, Gaussian random variables satisfy that $X + Y \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$ if $X \sim \mathcal{N}(0, \sigma_1^2), Y \sim \mathcal{N}(0, \sigma_2^2)$ are independent. Therefore, if each private data party adds *i.i.d.* Gaussian noise following $\mathcal{N}(0, \sigma^2)$ with $\sigma = 4\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon\sqrt{L}}$, the total noise follows $\mathcal{N}(0, \sigma_{total}^2)$ which guaranties (ϵ, δ)-DP for the whole WASP process with $\delta > \delta_{iter} \cdot (T - 1)$. □

E ADDITIONAL RESULTS

E.1 COMPARISON OF SYNTHETIC SAMPLE RESEMBLANCE FOR WASP

To further demonstrate the effectiveness of WASP under limited private sample setting, we additionally use FID between the generated DP synthetic dataset $\mathcal{D}$ and the real private dataset $\mathcal{B}$ to evaluate the resemblance of the former ($\mathcal{D}$) to the later ($\mathcal{B}$) with $M = 100$ in Figure 5. Lower FID indicates higher distribution similarity therefore indicating better resemblance.

As shown in Figure 5, the baseline method Aug-PE often fails to generate a DP synthetic dataset that closely resembles $\mathcal{B}$ when using an improper PLM, leading to an increased FID value over iterations. On the contrary, WASP results in a consistently decreasing FID value over iteration, demonstrating it effectiveness in improving the resemblance of $\mathcal{D}$ to $\mathcal{B}$. Moreover, although WASP initially has a higher FID than using the best single PLM (Flan-T5 in Figure 5) for Aug-PE (which is reasonable due

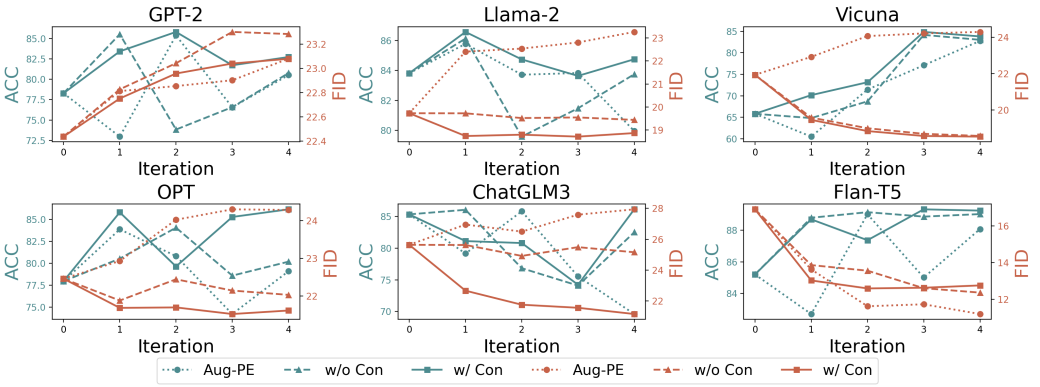


Figure 6: Comparison of the resemblance of synthetic dataset to real private dataset (Fréchet Inception Distance, FID) and trained downstream model performance (ACC) using Aug-PE ("Aug-PE", doted lines), refinement on $Q = 8$ without contrastive in-context learning ("w/o Con", dashed lines) and refinement on $Q = 8$ with contrastive in-context learning ("w/ Con", solid lines) with single-PLM setting and $L = 1$ under the same DP setting as in Table 1 with IMDb dataset.

to the initialization of $\mathcal{D}$ being a mixture of synthetic samples from different PLMs, making it better than the one generated solely by worst PLM but worse than the one given solely by the best PLM), it ultimately achieves a lower FID than all baseline counterparts. This indicates that our proposed WASP method better handles the limited private sample setting.

E.2 Effectiveness of Differentially Private Top-Q Voting and Contrast In-context Learning with Single PLM

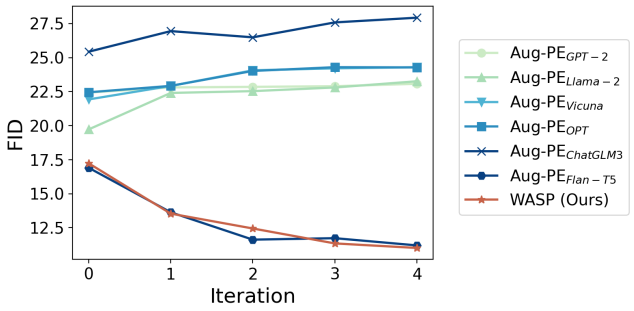


Figure 5: Comparison of the resemblance of synthetic dataset to real private dataset (FID) using Aug-PE and our proposed WASP using movie review semantic analysis task and IMDb dataset.

We present additional results to validate the effectiveness of *Differentially Private Top-Q voting* and *contrastive in-context learning* with single PLM in Figure 6. Starting with Aug-PE, we increase Q from 1 to 8 to obtain the "w/o Con" results, and then incorporate contrastive in-context learning samples into the prompt to achieve the "w/ Con" results (also the $K = 1$ setting for WASP). This refinement process shows a steady decrease in FID for most PLMs. Nonetheless, an overall performance improvement is observed for all tested PLMs, both in terms of highest performance across iterations and final performance.

E.3 Sensitive Analysis of M for PE

We performed experiments to analyze the sensitivity of Aug-PE Xie et al. (2024) on various M values. Results are included in Figure 7 which shows that most PLMs fail when only a limited amount of private samples ($M = 100$) is available, with an increasing FID through iterations. Conversely, with sufficient amount of private samples ($M = 10,000$), a continuous decrease in FID as well as less performance fluctuation can be observed throughout the iterations.

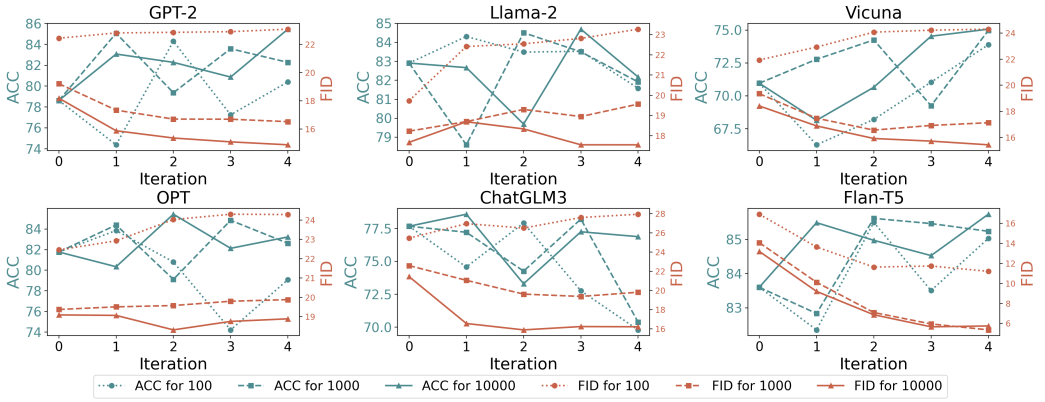


Figure 7: Comparison of the similarity of synthetic dataset to real private dataset (FID) and trained downstream model performance (ACC) with different amount of available private samples (M) using Aug-PE with $L = 1$ under the same DP setting as in Table 1 with IMDb dataset.

E.4 Comparison of WASP and Pre-Text Under User-level DP

In our work, we assume a full participation setting where all L parties participate in each iteration. Based on this, we primarily focus on ensuring sample-level DP to protect each private sample $(\mathbf{z}_j, u_j) \in \mathcal{B}$ in this work. However, our proposed WASP method can be easily extended to user-level DP protection and is also effective in protecting user-level DP compared to baselines (see Table 10).

In Hou et al. (2024), although they also study a full participation setting with $L > 1$, they focus on user-level DP with the assumption that each participating private data party in the collaboration controls only a tiny amount of private samples (8 in their work). Therefore, following Hou et al. (2024), to testify the effectiveness of WASP when extended to user-level DP, we assume that each participating data party controls no more than 8 real private samples, i.e. $M_l \leq 8, l = 1, \ldots, L$. These distributed private datasets still aggregate to an unbalanced dataset like in Section 5.1.

Under this setting, to protect user-level DP (where adding or removing one private data party should not significantly affect the function output), the function sensitively Δ_{user} should be $\max(M_1, \ldots, M_L)$ times as large as that for protecting sample-level DP (Δ_{sample}). The rationale is that, the addition or removal of a private data party can result in the addition or removal of up to $\max(M_1, \ldots, M_L)$ samples, leading to a change of no more than $\max(M_1, \ldots, M_L) \times \Delta_{sample}$ in the ℓ_2 distance of the produced histograms. Given that $\Delta_{sample} = 4$ for WASP (see Theorem D.2 in Appendix D for details), we have $\Delta_{user} = \max(M_1, \ldots, M_L) \times \Delta_{sample} \leq 8 \times \Delta_{sample} = 8 \times 4 = 32$. In our experiments, we use 32, the upper bound of Δ_{user}, as the function sensitivity to calculate $\sigma = 32\frac{\sqrt{2\log(1.25/\delta_{iter})}\sqrt{T-1}}{\epsilon\sqrt{L}}$ for (ϵ, δ)-DP protection with $\delta > \delta_{iter} \cdot (T-1)$.

Results are shown in Table 10 with a total of $L = 150$ private data parties controlling $M = 500$ private samples in total. Other experimental settings are the same with those in Table 2. Results show that, WASP continues to outperform baseline methods, including Pre-Text. This demonstrates that WASP is effectiveness not only under the need of guaranteeing sample-level DP but also under the need of providing user-level DP protection compared to baseline methods.

E.5 Comparison of Communication Overhead of WASP and Pre-Text for Federated Data Setting

We compare the transmitted information for secure aggregation between the baseline method Pre-Text and our proposed WASP framework in Table 11. With the same number of participating data parties (L), WASP only requires aggregating additional L histograms of dimension $\mathbb{R}^{|\mathcal{D}|}$ and uploading the aggregated histogram $H^f \in \mathbb{R}^{|\mathcal{D}|}$. These additional communicated information leads to only a minor increase in communication overhead compared to Pre-Text.

Table 10: Evaluation of downstream STM accuracy using 6 PLMs, $L = 150$. User-level DP is guaranteed instead of sample-level DP in this table. **Best** and second best results are marked.

		Privacy	$\lvert\mathcal{B}\rvert$	$\lvert\mathcal{D}\rvert$	IMDb	Yelp		Openreview		Banking
						Category	Rating	Area	Rating	
OnlyPrivate		$\epsilon = \infty$	500	-	83.61	57.27	44.15	22.76	32.79	74.56
FuseGen		Absolutely Private	-	6,000	89.07	63.38	57.96	24.70	34.57	78.75
Pre-Text	GPT-2	$\epsilon = 4.0$	500	6,000	83.96	63.04	45.78	27.46	24.09	75.75
	Llama-2	$\epsilon = 4.0$	500	6,000	84.28	60.24	50.54	29.02	34.15	82.50
	Vicuna	$\epsilon = 4.0$	500	6,000	83.67	63.21	51.42	28.18	32.87	83.38
	OPT	$\epsilon = 4.0$	500	6,000	84.69	62.92	50.40	28.59	24.29	81.25
	ChatGLM3	$\epsilon = 4.0$	500	6,000	85.56	57.46	51.54	29.78	32.33	84.88
	Flan-T5	$\epsilon = 4.0$	500	6,000	88.71	58.46	58.37	29.81	34.02	74.13
WASP (Ours)		$\epsilon = 4.0$	500	6,000	**89.15**	**63.49**	**59.78**	**29.96**	**37.10**	**85.25**

Table 11: Comparison of the information data parties' download, internal exchange and update in Pre-Text and WASP.

	Download	Exchange	Upload
Pre-Text	embedding of each $(\mathbf{x}_i, y_i) \in \mathcal{D}$	$\{H_l^n\}_{l=1}^L$	H^n
WASP (Ours)	embedding of each $(\mathbf{x}_i, y_i) \in \mathcal{D}$	$\{H_l^n\}_{l=1}^L, \{H_l^f\}_{l=1}^L$	H^n, H^f

E.6 More Strict DP Guarantee

In previous experiments, we use $\delta_{iter} = 1 \times 10^{-5}$ which will result in $\delta > 4 \times 10^{-5}$ for the whole process for all PE series baselines and WASP. Therefore, we compare using $\delta_{iter} = 1 \times 10^{-5}$ with $\delta = 1 \times 10^{-5}$ in Table 12. With $\delta = 1 \times 10^{-5}$, following Kairouz et al. (2015), $\delta_{iter} = 1 \times 10^{-23}$ can be applied to guarantee overall $(4, 1 \times 10^{-5})$-DP, which results in a noise scale roughly 2.14 times as large as the original one used in our original experiments in the paper.

These results demonstrate that, under tighter privacy guarantee ($\delta_{iter} = 1 \times 10^{-23}$, i.e. $\delta = 1 \times 10^{-5}$), the performance decrease is just minor, indicating the robustness of WASP and PE baselines.

F Additional Related Works

Due to space limitation, we include the discussion of previous works related to Contrastive In-context Learning (Contrastive ICL) here in the Appendix.

Contrastive In-context Learning.[1] The idea of using contrastive information to enrich in-context learning samples has been exploited from different aspects. Samples belonging to positive and negative classes Liang et al. (2024), correct or wrong self-predictions of training samples during training time Mo et al. (2024), human-preferred and non-preferred question responses Gao & Das

[1]Works Ren & Liu (2024); Miyanishi & Nguyen (2024) considering understanding in-context learning with contrastive learning theories are sometimes referred to using the same name, but we do not consider them here.

Table 12: Comparison of different DP δ using 6 open-source PLMs for PE baseline and our proposed WASP with $L = 1, M = 100$.

		Aug-PE						WASP
	GPT-2	Llama-2	Vicuna	OPT	ChatGLM3	Flan-T5		(Ours)
IMDb	$\delta_{iter} = 1 \times 10^{-5}$	85.38	85.77	82.76	83.86	85.82	89.00	89.52
	$\delta = 1 \times 10^{-5}$	84.88	85.30	82.04	83.52	85.22	88.83	89.18
Yelp	$\delta_{iter} = 1 \times 10^{-5}$	45.28	47.42	54.42	50.81	55.17	58.69	61.21
	$\delta = 1 \times 10^{-5}$	45.03	47.10	54.09	50.47	54.97	58.61	61.05

(2024) have all been utilized as contrastive samples. Our study is the first known effort to consider contrastive in-context learning for synthetic data generation, by treating synthetic samples of different qualities generated by multiple PLMs as contrastive information.

Algorithm 3 Functions used in Algorithms 1 and 2 for WASP

function `WeightedSynDataGeneration`($\mathcal{D}_k$, $\hat{\mathcal{D}}^n$, $\hat{\mathcal{D}}^f$, $\hat{N}$, C):
  for $c = 1$ **to** C **do**
    if $t = 0$ **then**
      Use zero-shot prompt as working prompt $\mathcal{T}(c)$.
    else
      Randomly sample $S - \lfloor S/2 \rfloor$ samples from $\hat{\mathcal{D}}^{n,[c]}$ and $\lfloor S/2 \rfloor$ samples from $\hat{\mathcal{D}}^{f,[c]}$ to create few-shot prompt as working prompt $\mathcal{T}(c)$.
    end if
    Generate $\lceil \hat{N}/C \rceil$ samples using $\mathcal{T}$ and add them to $\mathcal{D}_k$.
  end for
  if $|\mathcal{D}_k| > \hat{N}$ **then**
    Random sample $|\mathcal{D}_k| - \hat{N}$ different samples from $\mathcal{D}_k$ and remove them from $\mathcal{D}_k$.
  end if
  return $\mathcal{D}_k$.

function `STMTraining`($\mathcal{D}$, $m_{(0)}$, η):
  Initialize a trainable STM $m \leftarrow m_{(0)}$.
  Train m using $\mathcal{D}$ with learning rate η till convergence by using objective function $\mathcal{L} = \sum_{i=1}^{|\mathcal{D}|} \ell(m(\mathbf{x}_i), y_i)$.
  return m.

function `DP_PrivateVoting`($\mathcal{D}$, $\mathcal{B}$, Q, σ):
  Initialized $H^n \leftarrow [0, \ldots, 0]$; $H^f \leftarrow [0, \ldots, 0]$ of length $\mathbb{R}^{|\mathcal{D}|}$ and note the total DP synthetic dataset as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$.
  for $(\mathbf{z}_j, u_j)$ **in** $\mathcal{B}$ **do**
    $\mathcal{D}^{[u_j]} = \{(\mathbf{x}_i, y_i) \in \mathcal{D} \,|\, y_i = u_j\}$.
    $[n_{j,1}, \ldots, n_{j,Q}] \leftarrow \arg\text{top}Q\text{Smallest}\left(d(\mathbf{z}_j, \mathbf{x}_i)_{(\mathbf{x}_i, y_i) \in \mathcal{D}^{[u_j]}} \right)$.
    $[f_{j,1}, \ldots, f_{j,Q}] \leftarrow \arg\text{top}Q\text{Largest}\left(d(\mathbf{z}_j, \mathbf{x}_i)_{(\mathbf{x}_i, y_i) \in \mathcal{D}^{[u_j]}} \right)$.
    for $q = 1$ **to** Q **do**
      $H^n[n_{j,q}] \leftarrow H^n[n_{j,q}] + \frac{1}{2^{q-1}}$, $H^f[f_{j,q}] \leftarrow H^f[f_{j,q}] + \frac{1}{2^{q-1}}$.
    end for
  end for
  $H^n \leftarrow H^n + \mathcal{N}(0, \sigma^2 I_{|\mathcal{D}|})$, $H^f \leftarrow H^f + \mathcal{N}(0, \sigma^2 I_{|\mathcal{D}|})$.
  return H^n, H^f.

function `PLMScoring`(H, $\{\mathcal{D}_k\}_{k=1}^K$):
  for $k = 1$ **to** K **do**
    Calculate $s_i = H^n[i] \,/\, \sum_{i'=1}^{|\mathcal{D}|} H^n[i']$.
    Calculate model score $w_k = \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} s_i}{|\mathcal{D}_k| / \sum_{k'=1}^K |\mathcal{D}_{k'}|} = \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} s_i}{|\mathcal{D}_k|/|\mathcal{D}|}$
  end for
  return $\{w_k\}_{k=1}^K$.

function `SampleSelection`($\mathcal{D}$, H^n, H^f, S, C):
  Reset $\hat{\mathcal{D}}^n \leftarrow \emptyset$, $\hat{\mathcal{D}}^f \leftarrow \emptyset$.
  $\mathcal{H}^n \leftarrow H^n / \sum_{i=1}^{|\mathcal{D}|} H^n[i]$, $\mathcal{H}^f \leftarrow H^f / \sum_{i=1}^{|\mathcal{D}|} H^f[i]$.
  for $c = 1$ **to** C **do**
    $\mathcal{D}^{[c]} = \{(\mathbf{x}_i, y_i) \in \mathcal{D} \,|\, y_i = c\}$.
    $H^{n,[c]} = \left\{ H^n[i] \,\middle|\, (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \right\}$, $H^{f,[c]} = \left\{ H^f[i] \,\middle|\, (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \right\}$.
    $\hat{\mathcal{D}}^{n,[c]} = \left\{ (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \,\middle|\, H^n[i] \text{ is among the top-}S \text{ values of } H^{n,[c]} \right\}$, contains the top-$S$ samples from $\mathcal{D}^{[c]}$ ranked by $\mathcal{H}^{n,[c]}$.
    $\hat{\mathcal{D}}^{f,[c]} = \left\{ (\mathbf{x}_i, y_i) \in \mathcal{D}^{[c]} \,\middle|\, H^f[i] \text{ is among the top-}S \text{ values of } H^{f,[c]} \right\}$, contains the top-$S$ samples from $\mathcal{D}^{[c]}$ ranked by $\mathcal{H}^{f,[c]}$.
  end for
  $\hat{\mathcal{D}}^n = \left\{ \hat{\mathcal{D}}^{n,[1]}, \ldots, \hat{\mathcal{D}}^{n,[C]} \right\}$, $\hat{\mathcal{D}}^f = \left\{ \hat{\mathcal{D}}^{f,[1]}, \ldots, \hat{\mathcal{D}}^{f,[C]} \right\}$.
  return $\hat{\mathcal{D}}^n, \hat{\mathcal{D}}^f$.