MAPLE: Metadata Augmented Private Language Evolution

Eli Chien[*]
National Taiwan University
elichien@ntu.edu.tw

Yuzheng Hu[*] [†]
University of Illinois Urbana-Champaign
yh46@illinois.edu

Ryan McKenna
Google Research
mckennar@google.com

Shanshan Wu
Google Research
shanshanw@google.com

Zheng Xu[‡]
Meta
zhx@meta.com

Peter Kairouz
Google Research
kairouz@google.com

Abstract

Differentially private (DP) fine-tuning of large language models (LLMs) requires massive compute and full model access, which rules out state-of-the-art proprietary APIs for general users. Generating DP synthetic data offers a practical workaround. This approach also allows for transparent exploratory data analysis and arbitrary reuse across downstream tasks, sidestepping the rigid constraints of a model's parameter space. Private Evolution (PE) provides a promising API-based framework for generating this data, but its success relies heavily on initialization. If the private data distribution falls too far outside the foundation model's pre-training priors – a common issue in highly specialized domain – PE struggles to align with the target data. This misalignment causes poor convergence, degraded utility, and wasted API calls. To solve this initialization bottleneck, we introduce Metadata Augmented Private Language Evolution (MAPLE). MAPLE extracts DP tabular metadata and uses in-context learning to firmly ground the initial synthetic distribution in the target domain. Our evaluations on domain-specific text generation tasks show that MAPLE yields a strictly better privacy-utility trade-off, converges significantly faster, and sharply reduces API costs compared to baseline PE methods. Our code is available at https://github.com/elichien-google/MAPLE.

1 Introduction

The remarkable capabilities of modern artificial intelligence, particularly Large Language Models (LLMs), are driven by training on vast quantities of user-generated text (Ouyang et al., 2022; Touvron et al., 2023; Liang et al., 2023). Such data, ranging from mobile keyboard inputs (Xu et al., 2023b; Zhang et al., 2025) to sensitive medical records (Rumshisky et al., 2016) and recommendation histories (Karatzoglou & Hidasi, 2017; Zhang et al., 2019), form the backbone of model performance. However, this reliance on large-scale, user-generated data necessitates rigorous privacy safeguards; LLMs are known to memorize their training data, posing significant risks of sensitive information leakage (Carlini et al., 2021; Lukas et al., 2023; Wang et al., 2023).

Differentially Private (DP) synthetic text has emerged as a compelling approach to mitigating these risks. Due to the post-processing property of DP (Dwork et al., 2014), synthetic datasets can be reused across diverse downstream tasks without incurring additional privacy costs if the training data for generating DP synthetic data are not used again in downstream tasks. Furthermore, sharing synthetic data enables practitioners to perform transparent exploratory data analysis without the opaque constraints of a model's parameter space. Consequently, generating high-utility DP synthetic data has become a central goal, often

[*]Equal contribution.
[†]Work done while interning at Google Research.
[‡]Work done at Google Research.

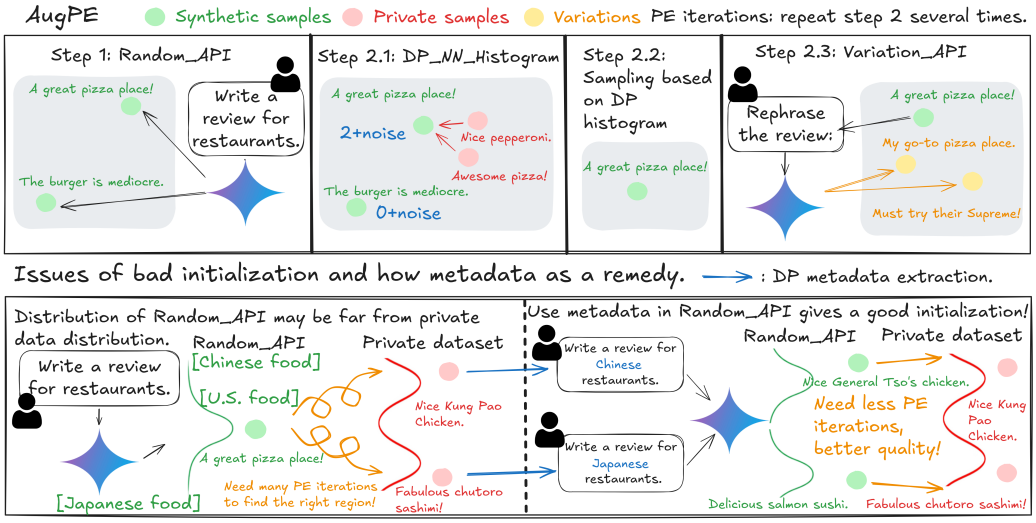

Figure 1: Illustration of AugPE and its limitations. **Top**: Overview of AugPE, adapted from Xie et al. (2024). AugPE first uses RANDOM_API to generate synthetic samples with a *data-independent* prompt (Step 1). It then iteratively refines these synthetic samples toward the private samples. At each iteration, AugPE selects synthetic samples that receive high nearest-neighbor votes from private samples using a differentially private voting histogram in an embedding space (Step 2.1 and 2.2). The selected synthetic samples are subsequently paraphrased via VARIATION_API to produce new synthetic samples (Step 2.3). This process is repeated for multiple rounds. **Bottom**: When the initial distribution induced by RANDOM_API is poorly aligned with the private samples, AugPE may require many iterations to reach the region of the private data distribution. To address this limitation, we propose incorporating metadata into the RANDOM_API prompt, making the initialization *data-dependent* and better aligned for the PE process. Details on differentially private metadata extraction are in Appendix A.2 and Figure 2.

described as the "holy grail" of privacy-preserving machine learning (Hu et al., 2024; Lin et al., 2024).

The most direct approach to generating DP synthetic text is to DP-finetune a generative language model on private corpora (Yue et al., 2023; Kurakin et al., 2023; Yu et al., 2024). However, several practical constraints often render this approach prohibitive. First, many state-of-the-art LLMs, such as Gemini 3 (Google Gemini Team, 2025), GPT 5.2 (OpenAI, 2025), (Anthropic, 2026), are accessible only through proprietary APIs. While standard finetuning APIs are available for a limited set of models (OpenAI, 2023), DP-finetuning requires specialized gradient-level updates (per-sample clipping and noise injection) that are not supported by any commercial API provider to date.

Second, even when high-performance open-source models like Llama (Grattafiori et al., 2024) or Qwen (Team, 2024) are available, DP-finetuning remains computationally expensive and challenging to implement (Sinha et al., 2025). This burden is particularly acute for data holders in resource-constrained settings, including medical clinics, financial institutions, and individual end-users. Moreover, similar difficulties arise even in large-scale industrial environments when private data must be processed within Trusted Execution Environments (TEEs) (Sabt et al., 2015) or under the constraints of federated learning (Kairouz et al., 2021), where access to system resources and gradient-level operations is inherently restricted.

Recently, the Private Evolution (PE) framework (Lin et al., 2024) has emerged as a promising alternative in settings where DP-finetuning is prohibitive. PE requires only API access to foundation models while achieving a privacy-utility trade-off competitive with state-of-the-art methods (Xie et al., 2024). However, the effectiveness of PE critically depends on the quality of its initialization, which is obtained via direct prompting (Figure 1). In particular, PE implicitly assumes that the initial unconditional generations are reasonably aligned with

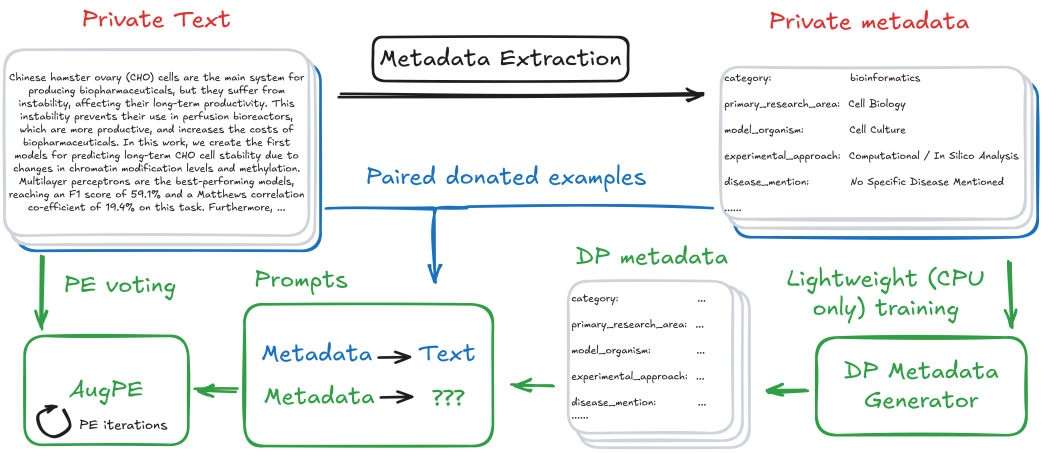

Figure 2: Overview of MAPLE. We first extract metadata in tabular format, either based on a designed schema or adapt it from the dataset whenever it is given. Then we train a light-weighted (with CPU only) DP metadata generator via the state-of-the-art approach, such as AIM (McKenna et al., 2022). Next, we compose the prompt for `RANDOM_API` with both DP metadata and a few donated (metadata, text) pairs as in-context examples. As we demonstrate in the ablation study in Section 5, our in-context example design is crucial for fully leveraging the in-context learning capability of LLMs. After the initialization, we refine it by the PE (i.e., AugPE) for the final DP synthetic dataset.

the target private data distribution. This assumption can break down when the private data distribution deviates substantially from the LLM's pre-training prior. These observations naturally raise a central research question: *"How can we design effective initializations for PE that improve the quality of DP synthetic data, especially under distributional mismatch?"*

We address this challenge by proposing **Metadata Augmented Private Language Evolution (MAPLE)**, a framework that leverages structural metadata to guide the evolution process while maintaining end-to-end differential privacy guarantees. A key feature of MAPLE is that it does not rely on pre-defined metadata being available in the raw private data; instead, we design a differentially private pipeline that extracts and synthesizes tabular metadata directly from the raw text. To further improve synthetic data quality, MAPLE exploits the in-context learning capabilities of LLMs by incorporating a small number of in-context examples into the prompt. Extensive experiments on two domain-specific datasets demonstrate that MAPLE consistently outperforms AugPE (Xie et al., 2024), the current state-of-the-art PE-based method, across all privacy regimes and evaluation metrics. Notably, MAPLE converges using substantially fewer API calls than AugPE. This reduction in query complexity leads to a more cost-effective and economically efficient solution for generating high-quality DP synthetic data, without requiring model finetuning.

2 Preliminaries

Differential Privacy (DP). DP (Dwork et al., 2006) is the gold standard for privacy, which is not only adopted by the U.S. government (Abowd, 2018) but also by the industry (Ding et al., 2017; Thakurta et al., 2017; Xu et al., 2023a). It provides a rigorous guarantee that limits what an adversary can infer about any single user's data from an algorithm's output.

Definition 2.1. A randomized algorithm $\mathcal{M}$ is (ε, δ)-DP if for any two adjacent datasets $\mathcal{D}, \mathcal{D}'$ (i.e., adding or removing one sample) and any subset of possible output S, we have

$$\mathbb{P}\left[\mathcal{M}(\mathcal{D}) \in S\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}(\mathcal{D}') \in S\right] + \delta. \tag{1}$$

$\varepsilon \in [0, \infty)$ is often referred to as the privacy loss, where a smaller ε implies a stronger privacy guarantee. δ represents a small failure probability, which is often set to be the reciprocal of the dataset size.

DP has many desirable properties for AI applications, especially private synthetic data generation. The *post-processing* property ensures that any post-processing of an (ε, δ)-DP mechanism does not incur additional privacy loss (Dwork et al., 2014) when the private dataset is not further touched. Moreover, the (sequential) composition theorem (Kairouz et al., 2015; Mironov, 2017) characterizes the overall privacy loss of the composition of multiple DP mechanisms.

Private Evolution (PE). PE (Lin et al., 2024), and its text-specific instantiation AugPE (Xie et al., 2024), are iterative algorithms that utilize two types of API calls to LLMs. The first is `RANDOM_API`, which generates an initial set of synthetic text with a data-independent prompt (Step 1)[1]. The second is `VARIATION_API`, which paraphrases a given text to produce a similar synthetic sample (Step 2.3). An overview of AugPE is shown in Figure 1. Notably, a key assumption underlying the effectiveness of PE-based methods is that the initial synthetic samples produced by `RANDOM_API` are reasonably aligned with the private data (Lin et al., 2024). Indeed, when most private data lie in the low probability region of the initial synthetic distribution, the initial samples can be far from the private data, causing PE to struggle to make progress. We illustrate this effect in the bottom row of Figure 1: repeatedly paraphrasing a review about "pizza" is unlikely to produce content related to Chinese or Japanese restaurants via `VARIATION_API`. As a result, a poorly aligned initialization can require many more PE iterations to achieve satisfactory synthetic data quality. This not only increases cost due to additional API calls, but also leads to a worse privacy-utility trade-off, since each iteration incurs additional privacy loss from calling `DP_NN_Histogram`.

3 MAPLE: Metadata Augmented Private Language Evolution

To address the poor initialization of PE, we propose to provide LLMs with additional information about the private dataset by incorporating metadata into the `RANDOM_API` prompt; see Figure 2 for an overview. The algorithm is provided in Appendix A.1.

Step 0: Pre-processing. We first extract metadata from the raw private text corpus. This step is crucial as it not only prevents us from assuming the existence of the auxiliary information in the private dataset, but also offers an opportunity to leverage domain knowledge for what metadata should be extracted. Following Hu et al. (2025), we choose to extract the metadata with a structured tabular schema. The schema is rich and multi-attribute with fixed options per attribute. The resulting metadata is in a tabular format and automatically annotated by LLMs. We refer interested readers to Appendix A.2 for further details on this pre-processing step.

Step 1: DP tabular metadata. While providing metadata in the prompt of `RANDOM_API` lets LLMs know more information about the private dataset, it also breaches the privacy of the private dataset. We need to ensure a proper DP guarantee for our metadata, which is also the reason why we restricted the metadata to the tabular format. Generating DP synthetic tabular data is a well-studied problem. In particular, the lightweight algorithm AIM (McKenna et al., 2022) is shown to offer strong tabular generation quality and only requires CPU training. This is particularly suitable for our resource-constrained scenario. We leverage AIM as our DP metadata generator, trained with private metadata.

Step 2: Prompt for `RANDOM_API`. To fully utilize the generation quality of powerful LLMs, we further utilize their in-context learning capability via few-shot examples. Before providing the synthetic metadata for text generation, we give a few (metadata, text) pairs as in-context examples. In practice, these non-private in-context examples can be either obtained from donated data from users who opt out of the privacy requirement or curated by the data holder (Nasr et al., 2023b; Wang et al., 2024). They are therefore outside the protected private dataset by construction. In our ablation study in Section 5, we found that providing such (metadata, text) pairs as in-context examples is crucial. Utilizing merely metadata or example text only provides a marginal improvement over the vanilla AugPE baseline.

[1]While (Xie et al., 2024) utilizes class labels available in the OpenReview dataset, they treat this information as public and do not ensure corresponding DP protection.

Step 3: AugPE. After obtaining the initial DP synthetic text dataset, we further refine it with a few AugPE iterations. As we will see in the experiment, starting with our enhanced initial DP synthetic text dataset not only results in better final synthetic data quality, but also converges with many fewer iterations. Interestingly, we find that it is still crucial to utilize a few AugPE iterations for the best overall results.

Privacy guarantee. MAPLE protects the private training corpus through the composition of two DP mechanisms: AIM for synthetic metadata generation and the Gaussian mechanism applied to the PE NN histograms. The LLM calls use only DP synthetic metadata, non-private in-context examples, or intermediate synthetic text, and are therefore post-processing of the DP mechanisms. Hence, the final synthetic text dataset satisfies an end-to-end DP guarantee with respect to all private training records.

4 Related Works

DP synthetic text. Methods for DP synthetic text generation are primarily divided into two paradigms: DP-finetuning and Private Evolution (PE). DP-finetuning approaches optimize pre-trained LLMs on private corpora using differentially private optimizers, most notably DP-SGD (Abadi et al., 2016; Yue et al., 2023; Carranza et al., 2024; Ochs & Habernal, 2025). Despite their effectiveness, these methods are often prohibitive when using closed-source models like GPT-4 or when operating within resource-constrained environments such as Trusted Execution Environments (TEEs) (Sabt et al., 2015).

In contrast, the PE paradigm (Lin et al., 2024) relies exclusively on LLM inference, making it better suited for black-box or compute-limited scenarios. Recent advancements in PE have focused on algorithmic refinements (Zou et al., 2025) and extensions to federated learning frameworks (Hou et al., 2024; 2025). These developments are orthogonal to our work; while prior research optimizes the iterative evolution process, MAPLE focuses on enhancing the initialization phase through DP-protected metadata and in-context learning.

Another emerging category is private inference, which injects DP noise directly into the next-token logits during the prompting process (Amin et al., 2024). However, private inference typically incurs higher privacy costs as the length of the generated sequence increases, and will feed private text to LLMs for obtaining logits. We view PE and private inference as complementary approaches with distinct use cases. Finally, our work bridges the gap between text and structured data by incorporating principles from DP tabular synthesis (Hu et al., 2024) into the text domain. For a more comprehensive overview of the broader DP synthetic data landscape, we refer readers to recent surveys by Hu et al. (2024) and Ponomareva et al. (2025).

Leveraging metadata. Recently, several works have been proposed to improve DP synthetic text generation via the use of metadata. Tan et al. (2025); Yu et al. (2024) improves the DP finetuning-based approach by distribution alignment defined on metadata in an *a priori* and *post-hoc*, respectively. Hu et al. (2025) further extends the idea of Tan et al. (2025) by introducing RL finetuning to have a better controllability for conditional generation. Very recently, Sun et al. (2025) also proposed to leverage the idea of private metadata as part of their improved design for PE. However, they do not explore the in-context learning capability of LLMs as we do, but rather provide it directly in the prompts of API calls. As we have observed in our ablation study, providing both metadata and in-context examples is crucial for the best results.

5 Experiments

Datasets. We conduct experiments on two challenging domain-specific datasets. *bioRxiv* (Hou et al., 2025) consists of a corpus of 29k scientific abstracts from preprints, with an average length of about 300 tokens. *OpenReview* (Xie et al., 2024) contains roughly 8.4k peer reviews for ICLR 2023 submissions collected from the OpenReview website [2].

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

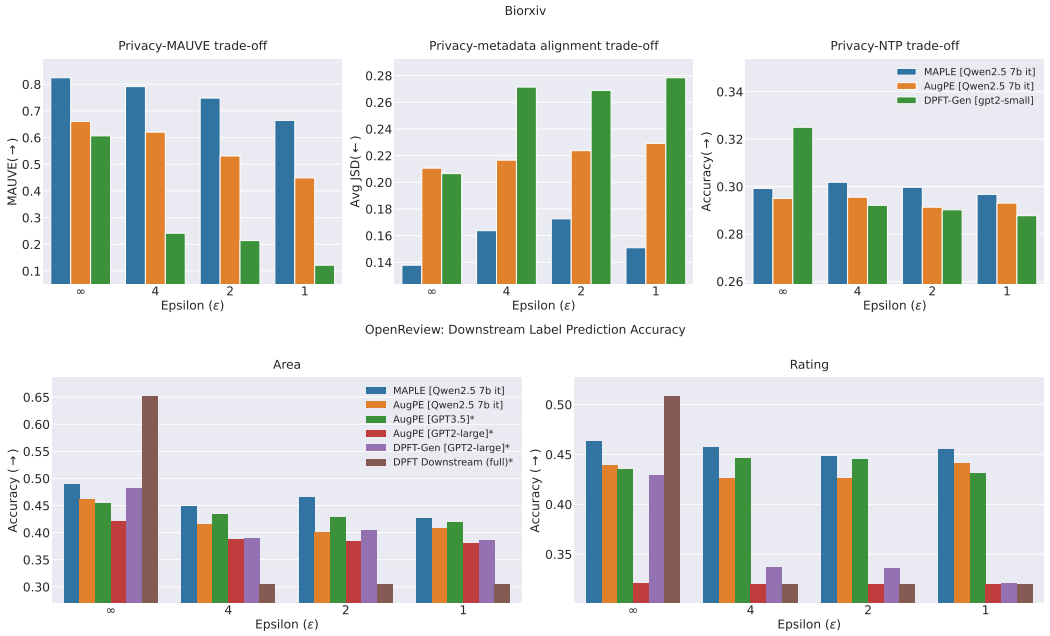

Figure 3: Main results. Top: Biorxiv datasets. The utility metrics are MAUVE score, average JSD on all metadata annotated by a powerful LLM, and NTP accuracy for training a bert-small model using synthetic data. Bottom: OpenReview datasets. The utility metrics are downstream prediction accuracy for using synthetic data to train a roberta-base model following the same setting of Xie et al. (2024), where the predicted labels are the area and rating of the review. The mark $*$ indicates the results are directly taken from Xie et al. (2024).

Compared to commonly used benchmarks such as Yelp Yelp (2025) in prior works (Yue et al., 2023; Xie et al., 2024; Tan et al., 2025), these datasets pose a more challenging synthetic text generation task. They are not only domain-specific, but also have longer average lengths (for example, Yelp has an average of 173 tokens). We provide further details for all datasets in Appendix B, and include an additional experiment on IMDB in Appendix C.1.

Evaluation metrics. We evaluate the synthetic data quality from multiple facets, capturing various aspects of the synthetic text. For *bioRxiv*, we assess the general fidelity by MAUVE (Pillutla et al., 2021) as in prior works (Xie et al., 2024; Hu et al., 2025). It quantifies the semantic similarity between the synthetic and private text at the distribution level. We chose sentence-t5-base (Ni et al., 2022) as the embedding model for MAUVE calculation, which better captures the text semantics compared to weaker embedding models such as stsb-roberta-base-v2 (Reimers & Gurevych, 2019) or all-MiniLM-L6-v2[3]. We also examine how the distribution of the annotated metadata from synthetic text aligns with the private one via the Jensen-Shannon distance between them. It provides a more fine-grained assessment of the particular aspects of the synthetic text compared to the general semantic meaning provided by MAUVE. For utility, we report the next token prediction (NTP) accuracy on a downstream generation task, where we trained a downstream GPT-2 small model (Radford et al., 2019) on the synthetic text and evaluated the NTP accuracy on the test set. For *OpenReview*, we follow the same setting as in Xie et al. (2024) to evaluate the downstream classification accuracy, where the labels are review recommendations (5 classes) and areas (12 classes). We finetune the RoBERTa-base (Liu et al., 2019) as text classifiers. Notably, prior work (Xie et al., 2024) treats these downstream labels as *public* and uses them from conditional generation in AugPE directly. In contrast, our MAPLE framework treats these labels as *private* metadata, which better aligns with the common practice.

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Baselines. We mainly compare with two baselines: AugPE (Xie et al., 2024) and DP finetuned generator (DPFT-Gen) (Yue et al., 2023). The setup of these baselines follows the implementation of Xie et al. (2024). For OpenReview, we directly take the results from Xie et al. (2024). For the other datasets that are not tested in Xie et al. (2024), we mainly follow their hyperparameter setting and optimized hyperparameters whenever necessary. Note that while there are some recent works that improve AugPE iterations (Zou et al., 2025; Sun et al., 2025), our improvement on initialization is orthogonal to them and may be combined for even better results. Since we focus solely on demonstrating the effect of better initialization for AugPE via metadata, we choose the vanilla AugPE as our primary baseline.

Implementation details. We choose `Qwen2.5-7B-Instruct`[4] (Team, 2024) as the base LLMs in our implementation of AugPE and MAPLE. Notably, the official AugPE implementation [5] uses the HuggingFace library as its backbone, which is sub-optimal in efficiency compared to the recent efficient LLM serving library such as vLLM (Kwon et al., 2023). This effect becomes even more critical in our MAPLE framework, as the in-context examples we provide in the prompt can have more than 3000 tokens. As a result, we re-implement AugPE with vLLM and use it as the building block of our MAPLE framework. This allows us to have a rigorous ablation study on the effect of initialization compared to taking the official implementation directly.

For all datasets, we fix the size of the donated text set to be 50, which is sampled from the validation set. Note that compared to both the targeted size of the synthetic dataset (2k or 5k), and the private training set size, assuming 50 donated examples is a practical setting (less than 3% at most). For each synthetic metadata, we choose the top 10 donated texts with the smallest Hamming distance on the metadata. We found that providing more than 10 in-context examples results in diminished return. For the metadata extraction, we utilize `gemini-2.5-flash-lite`[6] as the metadata annotator. Further discussion is deferred to Appendix A.

Privacy accounting. We evaluate all methods under privacy budgets $\varepsilon \in \{1, 2, 4, \infty\}$ with $\delta = 1/n$, where n is the private training set size. The total privacy cost of MAPLE comprises the DP metadata generator (AIM) and the PE voting iterations. We allocate the budget via zero-concentrated DP (zCDP) accounting (Bun & Steinke, 2016): the total (ε, δ) budget is converted to a zCDP parameter ρ, split as $\rho_{\text{AIM}} = 0.1\rho$ and $\rho_{\text{PE}} = 0.9\rho$. Each of the T voting iterations adds independent Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the nearest-neighbor voting histogram. By zCDP composition, the per-iteration cost is ρ_{PE}/T, yielding $\sigma = \sqrt{T/(2\rho_{\text{PE}})}$. For AugPE, the entire budget is allocated to PE voting.

Hyperparameters and hardware. The main hyperparameter is the number of private evolution iterations T. For both MAPLE and AugPE, we first run the non-private version with $T = 10$ to identify the iteration at which MAUVE peaks (Figure 5), then search locally around that value in the private setting. Unlike the non-private case where a single run with large T suffices, each choice of T under DP requires a separate run with a different noise multiplier σ, making exhaustive search over T computationally expensive. Following common practice in DP synthetic data generation (Xie et al., 2024), we report single-run results. All experiments were conducted on NVIDIA A100 40GB GPUs.

5.1 Main results

We first compare the end-to-end performance of MAPLE with the baselines on different datasets. MAPLE consistently outperforms AugPE across all datasets, metrics, and privacy constraints. It validates the effectiveness of leveraging metadata and in-context learning capability in MAPLE. We also observe that MAPLE can outperform the DP finetuning generator approach. This observation is aligned with Xie et al. (2024), where the PE-based approach may offer a better privacy-utility trade-off than the DP finetuned generator. Surprisingly, we observe that MAPLE with Qwen 2.5 7b can even surpass the performance

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[5] https://github.com/microsoft/DPSDA
[6] https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash-lite

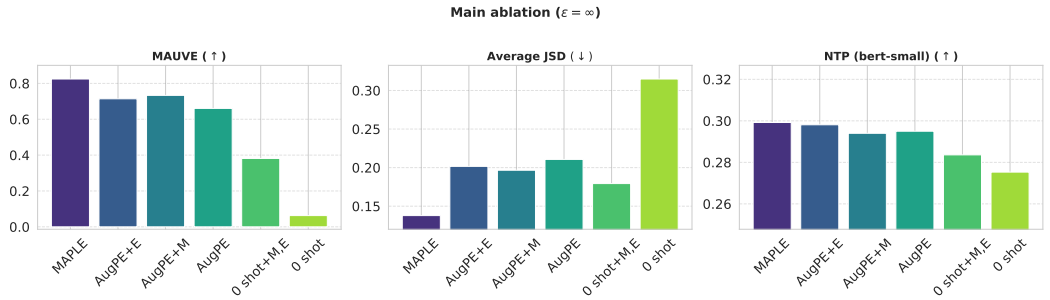

Figure 4: Ablation study on Biorxiv dataset. +M: leveraging only metadata in the prompt. +E: leveraging only in-context examples in the prompt. MAPLE: leverage both in the prompt. 0 shot: direct prompting without PE iterations.

of AugPE with GPT3.5 on the Openreview dataset, despite GPT3.5 being a much larger LLM. This result further highlights the importance of our design in MAPLE.

5.2 Ablation studies

MAPLE core components. We conduct a series of controlled ablation studies to isolate the contributions of different components in MAPLE. Specifically, we consider the following variants. (1) **AugPE+M**: only synthetic metadata is included in the `RANDOM_API` prompt. (2) **AugPE+E**: only 10 uniformly sampled donated examples are included in the prompt (note metadata is unavailable for similarity-based selection in this setting.) Recall that MAPLE leverages both metadata (M) and donated text (E), which can be viewed as **AugPE+M,E**. We additionally evaluate the quality of the initialization without PE iterations. (3) **0 shot**: the initial synthetic dataset in AugPE, without PE voting or iterative refinement. (4) **0 shot+M,E**: the MAPLE initialization, using both metadata and in-context examples, without PE voting or iterative refinement. The latter two ablation studies assess the importance of PE iterations and whether a good initialization alone is sufficient to produce high-quality synthetic data.

Figure 4 presents the comparison under $\varepsilon = \infty$; we observe that using either metadata or in-context examples alone yields limited improvement over AugPE. It is crucial to leverage both metadata and in-context examples to fully exploit LLM capabilities. We also confirm the necessity of PE iterations. While incorporating in-context examples in the 0-shot setting substantially improves generation quality, it still falls short of the performance achieved by MAPLE and even AugPE. These results show that PE iterations, metadata, and in-context examples are all essential for achieving high-quality synthetic data. The poor 0-shot results also suggest that general pretraining is insufficient to match the specific long-form distributions of bioRxiv abstracts without dataset-specific initialization.

This conclusion holds under finite privacy budgets: we compare MAPLE with AugPE+E on bioRxiv under $\varepsilon = 1$ and $\varepsilon = 4$. MAPLE achieves MAUVE scores of 0.6846 and 0.7388, respectively, compared with 0.4915 and 0.5732 for AugPE+E. This confirms that metadata contributes substantial gains beyond donated examples alone under varying DP constraints.

Metadata richness. We examine how the amount of information provided in the API prompt affects PE convergence, which can be viewed as an ablation on metadata richness (Figure 5). AugPE requires 8 iterations to reach its best generation quality. In contrast, MAPLE converges in only 2 iterations while achieving substantially higher quality. This not only reduces the cost of API calls, but also implies that less DP noise is needed for each PE iteration due to the DP composition theorem. Interestingly, we find that using weaker metadata with 2 attributes requires 4 iterations, more than the full metadata (9 attributes), but still far fewer than AugPE. These results indicate that richer metadata leads to faster PE convergence and improved peak performance. Under the DP budgets in Figure 3, MAPLE uses $T = 3, 3, 2$ PE iterations for $\varepsilon = 4, 2, 1$, while AugPE uses $T = 9, 9, 8$. Since other PE parameters are held fixed, this gives roughly a 3–4$\times$ **reduction** in API calls.

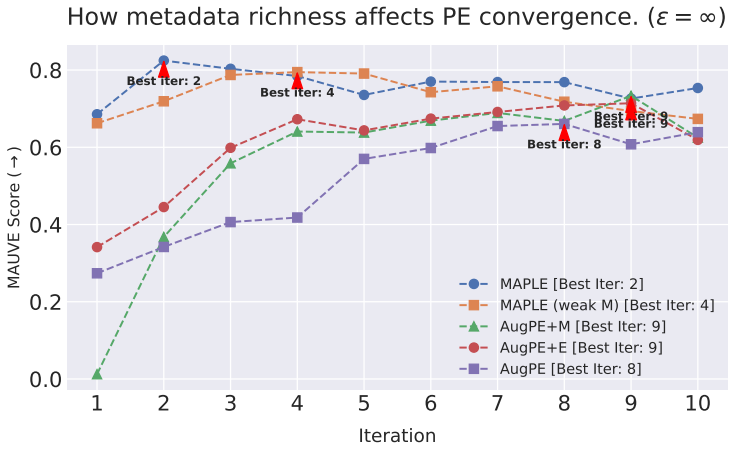

Figure 5: How metadata richness affects PE convergence, measured by MAUVE score on the Biorxiv dataset. Weak M uses only two metadata attributes from the full set.

6 Conclusion and Future Directions

In this work, we introduced MAPLE, a framework designed to overcome the initialization bottleneck in the Private Evolution (PE) algorithm. By synergistically leveraging DP-extracted tabular metadata and in-context examples, MAPLE provides a more informed starting point for synthetic data generation. Our extensive experiments and ablation studies underscore the critical role of initialization in API-based DP frameworks and demonstrate that MAPLE significantly improves convergence and utility while maintaining strict privacy guarantees.

Looking ahead, several promising avenues for future research remain:

- **Extending Beyond Initialization:** In its current form, MAPLE primarily focuses on optimizing the initial distribution (i.e., `RANDOM_API`). However, the insights derived from metadata and in-context learning could inherently benefit other modules within the PE pipeline, such as the `VARIATION_API`. While we maintained a controlled experimental setup to isolate the effects of initialization, investigating how these components can assist in the iterative evolution process could further enhance the privacy-utility trade-off.
- **Selection Strategy:** Existing PE implementations, starting from the original PE work, typically select the top-K samples from the DP voting histogram at each iteration. This simple strategy may reduce diversity when many similar samples receive high votes. Designing and testing diversity-aware selection strategies is an interesting direction for future work.
- **Cross-Modal Generalization:** A natural extension is to adapt the MAPLE framework to other data modalities, including images, tabular data, and multimedia (audio/video). Determining how to extract representative and helpful metadata under DP constraints for non-textual data remains a compelling challenge for the PE-based approach.
- **Fully Private Exemplar Selection:** Although our current implementation utilizes a small set of "donated" data points for in-context learning, this requirement can be further relaxed. When public or donated examples are unavailable, private-prediction-based algorithms such as Amin et al. (2024) could be used to "rewrite" private examples with a formal DP guarantee, at the cost of additional privacy budget. We expect minor shifts between donated and private examples to be less harmful when formatting, terminology, and style are preserved; systematically studying this sensitivity is a useful direction for future work.

- **A Unified Private Evolution Library:** Currently, there is a notable lack of comprehensive frameworks for developing new private evolution (PE) algorithms and corresponding benchmarking pipelines. While Lin et al. (2024) provide an excellent DPSDA library[7] for PE, it is built upon Hugging Face's transformers library and lacks direct support for efficient inference libraries like vLLM (Kwon et al., 2023). The computational speedup offered by vLLM over the standard transformers library is substantial; this efficiency gain was the primary motivation for re-implementing our AugPE method using vLLM. Integrating vLLM into the existing DPSDA ecosystem would be a valuable contribution to the community. Additionally, there remains a critical need for robust evaluation pipelines tailored to PE-based algorithms and, more broadly, differentially private (DP) synthetic data generation methods. Although recent efforts have made promising strides in this direction (Wang et al., 2025), the establishment of comprehensive, standardized benchmarks remains an important goal for the community.
- **Privacy Auditing for PE-based Methods:** While PE-based approaches offer rigorous DP guarantees that provide a theoretical upper bound on privacy loss for generated synthetic datasets, empirical privacy auditing remains crucial to establish a practical lower bound. This mirrors the well-established need for auditing in DP model training (Nasr et al., 2023a; Steinke et al., 2023; Cebere et al., 2025). Nevertheless, auditing methods tailored specifically to DP synthetic data generation remain significantly underexplored. Furthermore, the limited existing work on auditing synthetic data privacy (Meeus et al., 2025) is not directly applicable to PE-based methods. Addressing this methodological gap represents an important direction for future research within the community.

Acknowledgments

We thank Da Yu for the help with the Gemini API settings. We thank Zinan Lin and Chulin Xie for clarifying questions regarding AugPE. We thank Brendan McMahan and Daniel Ramage for valuable discussions and feedback on this work.

EC is supported by the Yushan Young Fellow Program (115V1070-1) from the Ministry of Education, Taiwan, and NTU Artificial Intelligence Center of Research Excellence (114M7069-01) within Taiwan Centers of Excellence in Artificial Intelligence.

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.

John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2867–2867, 2018.

Kareem Amin, Alex Bie, Weiwei Kong, Alexey Kurakin, Natalia Ponomareva, Umar Syed, Andreas Terzis, and Sergei Vassilvitskii. Private prediction for large-scale synthetic text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7244–7262, 2024.

Anthropic. Introducing claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, February 2026. Model release announcement.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pp. 635–658. Springer, 2016.

[7] https://github.com/microsoft/DPSDA

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

Aldo Carranza, Rezsa Farahani, Natalia Ponomareva, Alexey Kurakin, Matthew Jagielski, and Milad Nasr. Synthetic query generation for privacy-preserving deep retrieval systems using differentially private language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3920–3930, 2024.

Tudor Ioan Cebere, Aurélien Bellet, and Nicolas Papernot. Tighter privacy auditing of DP-SGD in the hidden state threat model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xzKFnsJIXL.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *Advances in Neural Information Processing Systems*, 36:76852–76871, 2023.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. 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.

Google Gemini Team. Gemini 3 pro model card. Technical report, Google DeepMind, November 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng LI, Bo Li, and Zhangyang Wang. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ifz3IgsEPX.

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. URL https://openreview.net/forum?id=3WCvnkHnxV.

Charlie Hou, Mei-Yu Wang, Yige Zhu, Daniel Lazar, and Giulia Fanti. Private federated learning using preference-optimized synthetic data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZuaU2bYzlc.

Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. Sok: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4696–4713. IEEE, 2024.

Yuzheng Hu, Ryan McKenna, Da Yu, Shanshan Wu, Han Zhao, Zheng Xu, and Peter Kairouz. Actg-arl: Differentially private conditional text generation with rl-boosted control. *arXiv preprint arXiv:2510.18232*, 2025.

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

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Alexandros Karatzoglou and Balázs Hidasi. Deep learning for recommender systems. In *Proceedings of the eleventh ACM conference on recommender systems*, pp. 396–397, 2017.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

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. URL https://openreview.net/forum?id=YEhQs8POIo.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE, 2023.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: an adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11):2599–2612, 2022.

Matthieu Meeus, Lukas Wutschitz, Santiago Zanella-Beguelin, Shruti Tople, and Reza Shokri. The canary's echo: Auditing privacy risks of LLM-generated synthetic text. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=f3mQ0xYA1I.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023a.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In *International Conference on Machine Learning*, pp. 25718–25732. PMLR, 2023b.

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.

Sebastian Ochs and Ivan Habernal. Private synthetic text generation with diffusion models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10612–10626, 2025.

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

OpenAI. Gpt-5.2 system card and release notes. https://help.openai.com/en/articles/6825453-chatgpt-release-notes, December 2025. See also: GPT-5 System Card (August 2025).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

Natalia Ponomareva, Zheng Xu, H Brendan McMahan, Peter Kairouz, Lucas Rosenblatt, Vincent Cohen-Addad, Cristóbal Guzmán, Ryan McKenna, Galen Andrew, Alex Bie, et al. How to dp-fy your data: A practical guide to generating synthetic data with differential privacy. *arXiv preprint arXiv:2512.03238*, 2025.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

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.

Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, volume 1, pp. 57–64. IEEE, 2015.

Amer Sinha, Thomas Mesnard, Ryan McKenna, Daogao Liu, Christopher A Choquette-Choo, Yangsibo Huang, Da Yu, George Kaissis, Zachary Charles, Ruibo Liu, et al. Vaultgemma: A differentially private gemma model. *arXiv preprint arXiv:2510.15001*, 2025.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36:49268–49280, 2023.

Zhonghao Sun, Zhiliang Tian, Yiping Song, Yuyi Si, Juhua Zhang, Minlie Huang, Kai Lu, Zeyu Xiong, Xinwang Liu, and Dongsheng Li. Dpga-textsyn: Differentially private genetic algorithm for synthetic text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16159–16179, 2025.

Bowen Tan, Zheng Xu, Eric P. Xing, Zhiting Hu, and Shanshan Wu. Synthesizing privacy-preserving text data via finetuning *without* finetuning billion-scale LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=FCm4laCLiH.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Abhradeep Guha Thakurta, Andrew H Vyrros, Umesh S Vaishampayan, Gaurav Kapoor, Julien Freudiger, Vivek Rangarajan Sridhar, and Doug Davidson. Learning new words. *Granted US Patents*, 9594741:2, 2017.

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. *arXiv preprint arXiv:2307.09288*, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

Ruizhe Wang, Roberta De Viti, Aarushi Dubey, and Elissa M Redmiles. The role of privacy guarantees in voluntary donation of private data for altruistic goals. *arXiv e-prints*, pp. arXiv–2407, 2024.

Shuaiqi Wang, Vikas Raunak, Arturs Backurs, Victor Reis, Pei Zhou, Sihao Chen, Longqi Yang, Zinan Lin, Sergey Yekhanin, and Giulia Fanti. Struct-bench: A benchmark for differentially private structured text generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=59vXWteYuh.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LWD7upg1ob.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 629–639, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.60. URL https://aclanthology.org/2023.acl-industry.60/.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 629–639, 2023b.

Yelp. Yelp open dataset, 2025. URL https://www.yelp.com/dataset.

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-preserving instructions for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 57480–57506, 2024.

Xiang Yue, Huseyin A 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 *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

Yanxiang Zhang, Zheng Xu, Shanshan Wu, Yuanbo Zhang, and Daniel Ramage. Synthesizing and adapting error correction data for mobile large language model applications. *arXiv preprint arXiv:2505.18488*, 2025.

Tianyuan Zou, Yang Liu, Peng Li, Yufei Xiong, Jianqing Zhang, Jingjing Liu, Xiaozhou Ye, Ye Ouyang, and Ya-Qin Zhang. Contrastive private data synthesis via weighted multi-plm fusion. *arXiv preprint arXiv:2502.00245*, 2025.

A Additional Details of MAPLE

A.1 MAPLE procedure

Inputs: private text corpus $\mathcal{D}_{\text{priv}}$, non-private donated exemplar set $\mathcal{D}_{\text{don}}$, privacy budget (ε, δ), target synthetic dataset size, and PE iteration count T.

Output: DP synthetic text dataset $\mathcal{D}_{\text{syn}}$.

1. Extract structured tabular metadata from each text in $\mathcal{D}_{\text{priv}}$.
2. Fit AIM on the private metadata and sample DP synthetic metadata using the allocated metadata budget.
3. For each synthetic metadata record, retrieve matching non-private exemplars from $\mathcal{D}_{\text{don}}$ and construct the `RANDOM_API` prompt.
4. Generate the initial synthetic text dataset conditioned on the DP synthetic metadata and non-private exemplars.
5. Run T AugPE refinement iterations, using DP nearest-neighbor voting followed by `VARIATION_API`.
6. Return the final synthetic text dataset $\mathcal{D}_{\text{syn}}$.

A.2 Metadata extraction

We follow the strategy of Hu et al. (2025) to extract metadata in the form of a structured tabular schema. We include the details here for self-completeness.

Schema design. For the biorxiv dataset, we follow the same schema as in Hu et al. (2025) with an additional word count field recording the word count of each text. For the OpenReview dataset, it is inherent with "area" and "recommendation" of the review. We additionally include the following information in the schema: primary contribution type, primary strength, primary weakness, novelty assessment, reviewer tone, clarity and presentation quality, actionability of feedback, word count, and three keywords. For the three keywords, they are chosen from a fixed, disjoint set of keyword pool, respectively. All of them have an option "Other", indicating that none of the keywords in the pool match the review. For the other fields, they all have their respective set of options. Full details can be found at Listing 1 and 2 for the biorxiv and OpenReview dataset respectively.

Metadata Extraction. We use `gemini-2.5-flash-lite` to perform metadata extraction. An example prompt for the biorxiv dataset can be at Listing 3. Note that we assume the data holder can locally host a metadata extractor that does not lead to a privacy breach in our experiment. In practice, if this assumption does not hold, the alternative solution is to leverage privacy-preserving inference methods (Duan et al., 2023; Hong et al., 2024) for secure API queries or manually annotate the metadata. While this might lead to lower-quality metadata, our ablation study (Figure 5) shows that even weak metadata still offers significantly better initialization than vanilla AugPE.

```
attribute_domains = {
  "primary_research_area": domain.CategoricalAttribute([
      "Biochemistry",
      "Bioinformatics",
      "Biophysics",
      "Cancer Biology",
      "Cell Biology",
      "Clinical Trials",
      "Developmental Biology",
      "Ecology",
      "Epidemiology",
      "Evolutionary Biology",
      "Genetics",
      "Genomics",
      "Immunology",
      "Microbiology",
      "Molecular Biology",
      "Neuroscience",
      "Paleontology",
```

```
        "Pathology",
        "Pharmacology and Toxicology",
        "Physiology",
        "Plant Biology",
        "Public Health",
        "Scientific Communication and Education",
        "Structural Biology",
        "Synthetic Biology",
        "Systems Biology",
        "Zoology",
        "Other",
    ]),
    "model_organism": domain.CategoricalAttribute([
        "Human",
        "Mouse/Rat",
        "Zebrafish",
        "Drosophila melanogaster",
        "Caenorhabditis elegans",
        "Saccharomyces cerevisiae",
        "Escherichia coli",
        "Arabidopsis thaliana",
        "Plant",
        "Cell Culture",
        "In Silico / Computational",
        "Other Mammal",
        "Other Vertebrate",
        "Other Invertebrate",
        "Other Microbe",
        "Not Applicable / Review",
        "Other",
    ]),
    "experimental_approach": domain.CategoricalAttribute([
        "Wet Lab Experimentation",
        "Computational / In Silico Analysis",
        "Clinical Study",
        "Field Study / Observation",
        "Case Study / Case Review",
        "Review / Meta-analysis",
        "New Method Development",
        "Theoretical Modeling",
        "Other",
    ]),
    "dominant_data_type": domain.CategoricalAttribute([
        "Genomic",
        "Transcriptomic",
        "Proteomic",
        "Metabolomic",
        "Imaging",
        "Structural",
        "Phenotypic / Behavioral",
        "Ecological / Environmental",
        "Clinical / Patient Data",
        "Simulation / Model Output",
        "Multi-omics",
        "Other",
    ]),
    "research_focus_scale": domain.CategoricalAttribute([
        "Molecular",
        "Cellular",
        "Circuit / Network",
        "Tissue / Organ",
        "Organismal",
        "Population",
        "Ecosystem",
        "Multi-scale",
        "Other",
    ]),
    "disease_mention": domain.CategoricalAttribute([
        "Cancer",
        "Neurodegenerative Disease",
        "Infectious Disease",
        "Metabolic Disease",
        "Cardiovascular Disease",
        "Autoimmune / Inflammatory Disease",
        "Psychiatric / Neurological Disorder",
        "Genetic Disorder",
        "No Specific Disease Mentioned",
        "Other",
    ]),
    "sample_size": domain.CategoricalAttribute([
        "Single Subject / Case Study",
```

```
        "Small Cohort (<50 subjects)",
        "Medium Cohort (50-1000 subjects)",
        "Large Cohort / Population-scale (>1000 subjects)",
        "Relies on Cell/Animal Replicates",
        "Not Specified / Not Applicable",
    ]),
    "research_goal": domain.CategoricalAttribute([
        "Investigating a mechanism",
        "Characterizing a system/molecule",
        "Developing a method/tool",
        "Identifying novel elements",
        "Testing a hypothesis",
        "Quantifying a parameter",
        "Evaluating/Comparing approaches",
        "Other",
    ]),
    "word_count": domain.CategoricalAttribute(np.arange(50.0,650.0,50)),
}
```

Listing 1: Metadata schema for biorxiv dataset

```
attribute_domains = {
    "area": domain.CategoricalAttribute([
        'Deep Learning and representational learning',
        'Applications (eg, speech processing, computer vision, NLP)',
        'Reinforcement Learning (eg, decision and control, planning, hierarchical RL,
      robotics)',
        'Neuroscience and Cognitive Science (e.g., neural coding, brain-computer interfaces)',
        'Probabilistic Methods (eg, variational inference, causal inference, Gaussian
      processes)',
        'Social Aspects of Machine Learning (eg, AI safety, fairness, privacy,
      interpretability, human-AI interaction, ethics)',
        'Unsupervised and Self-supervised learning',
        'Machine Learning for Sciences (eg biology, physics, health sciences, social sciences,
      climate/sustainability )',
        'General Machine Learning',
        'Theory (eg, control theory, learning theory, algorithmic game theory)',
        'Generative models',
        'Optimization (eg, convex and non-convex optimization)',
    ]),
    "recommendation": domain.CategoricalAttribute([
        '6: marginally above the acceptance threshold',
        '3: reject, not good enough',
        '5: marginally below the acceptance threshold',
        '8: accept, good paper',
        '1: strong reject'
    ]),
    "primary_strength": domain.CategoricalAttribute([
        'Novelty/Originality of Idea', 'Strong Empirical Performance', 'Significant Problem or
      Application',
        'High-Quality Presentation/Clarity', 'Solid Theoretical Guarantees', 'Good
      Reproducibility (Code/Data Provided)'
    ]),
    "primary_weakness": domain.CategoricalAttribute([
        'Insufficient/Flawed Experiments', 'Weak or Missing Baselines', 'Lack of Novelty',
        'Unclear/Poor Presentation', 'Flawed Theoretical Claims', 'Limited Scope or Impact',
      'Ethical Concerns'
    ]),
    "keywords_group_1": domain.CategoricalAttribute([
        'Large Language Models (LLMs)', 'Diffusion Models', 'Foundation Models', 'Prompting',
        'Text-to-Image Synthesis', 'Transformers', 'Self-Supervised Learning (SSL)',
        'Datasets & Benchmarks', 'Data Augmentation', 'Other'
    ]),
    "keywords_group_2": domain.CategoricalAttribute([
        'Few-Shot & Zero-Shot Learning', 'Meta-Learning', 'Continual Learning', 'Transfer
      Learning',
        'Federated Learning', 'Graph Neural Networks (GNNs)', 'Knowledge Graphs',
        'Weak Supervision', 'Causal Inference', 'Other'
    ]),
    "keywords_group_3": domain.CategoricalAttribute([
        'Efficiency', 'Stochastic Gradient Descent (SGD)', 'Model Compression', 'Robustness',
        'Out-of-Distribution (OOD) Generalization', 'Fairness & Bias', 'Privacy',
        'Interpretability & Explainability', 'Reinforcement Learning (RL)', 'Decision Making',
      'Other'
    ]),
    "word_count": domain.CategoricalAttribute(np.arange(50.0, 1200.0, 50)),
}
```

Listing 2: Metadata schema for OpenReview dataset

```
def annotate_features_prompt(
    text: str,
) -> str:
    """Returns a prompt to be used for feature extraction from text.

    This function generates a system prompt for an LLM to annotate a given text
    based on a predefined set of features. The features are derived from a
    Pydantic dataclass. The prompt instructs the LLM to output a JSON object
    containing the annotated features.

    Args:
    dataset_description: A 1-2 sentence description of the dataset.
    dataclass: A Pydantic dataclass defining the features to be extracted.
      Each field in the dataclass represents a feature, and the field's
      type annotation should specify the possible values (e.g., using
      typing.Literal).
    text: The text to be annotated.

    Returns:
    A detailed system prompt suitable for use with frontier language models
    for feature extraction, instructing the model to return a JSON object.
    """
    return textwrap.dedent("""\
      You are an expert biomedical information extraction assistant. Your task is to
     carefully read a scientific abstract from bioRxiv and extract the specified features
     according to the schema provided.

Output exactly one JSON object with no extra text or explanations.

**CRITICAL INSTRUCTION 1:** For any field in the JSON schema that lists specific options
     (e.g., "<Option1|Option2|...>"), you MUST select one of the provided options exactly as
     it is written. Do not invent, alter, or combine options. Failure to use an exact option
     from the list will be considered an error.

**CRITICAL INSTRUCTION 2:** Ensure the value chosen for a field is appropriate for that
     field's specific definition. Do not use an option from one field (e.g., 'Cellular' from
     `research_focus_scale`) as the value for another field.

Use this schema:
```json
{{
  "primary_research_area": "<Biochemistry|Bioinformatics|Biophysics|Cancer Biology|Cell
      Biology|Clinical Trials|Developmental Biology|Ecology|Epidemiology|Evolutionary
      Biology|Genetics|Genomics|Immunology|Microbiology|Molecular
      Biology|Neuroscience|Paleontology|Pathology|Pharmacology and Toxicology|Physiology|Plant
      Biology|Public Health|Scientific Communication and Education|Structural
      Biology|Synthetic Biology|Systems Biology|Zoology|Other>", // Categorizes the abstract
      into its main biological discipline.
  "model_organism": "<Human|Mouse/Rat|Zebrafish|Drosophila melanogaster|Caenorhabditis
      elegans|Saccharomyces cerevisiae|Escherichia coli|Arabidopsis thaliana|Plant|Cell
      Culture|In Silico / Computational|Other Mammal|Other Vertebrate|Other Invertebrate|Other
      Microbe|Not Applicable / Review|Other>", // Identifies the primary biological model used
      in the research.
  "experimental_approach": "<Wet Lab Experimentation|Computational / In Silico
      Analysis|Clinical Study|Field Study / Observation|Case Study / Case Review|Review /
      Meta-analysis|New Method Development|Theoretical Modeling|Other>", // Describes the main
      methodology used to conduct the study.
  "dominant_data_type":
      "<Genomic|Transcriptomic|Proteomic|Metabolomic|Imaging|Structural|Phenotypic /
      Behavioral|Ecological / Environmental|Clinical / Patient Data|Simulation / Model
      Output|Multi-omics|Other>", // Specifies the primary type of data generated or analyzed
      in the paper.
  "research_focus_scale": "<Molecular|Cellular|Circuit / Network|Tissue /
      Organ|Organismal|Population|Ecosystem|Multi-scale|Other>", // Categorizes the biological
      level of organization the study focuses on.
  "disease_mention": "<Cancer|Neurodegenerative Disease|Infectious Disease|Metabolic
      Disease|Cardiovascular Disease|Autoimmune / Inflammatory Disease|Psychiatric /
      Neurological Disorder|Genetic Disorder|No Specific Disease Mentioned|Other>", //
      Identifies whether the abstract explicitly names a disease or a major disease category.
  "sample_size": "<Single Subject / Case Study|Small Cohort (<50 subjects)|Medium Cohort
      (50-1000 subjects)|Large Cohort / Population-scale (>1000 subjects)|Relies on
      Cell/Animal Replicates|Not Specified / Not Applicable>", // Estimates the scale of the
      study based on mentions of sample or cohort size.
  "research_goal": "<Investigating a mechanism|Characterizing a system/molecule|Developing a
      method/tool|Identifying novel elements|Testing a hypothesis|Quantifying a
      parameter|Evaluating/Comparing approaches|Other>" // Categorizes the study's primary
      objective based on its framing.
}}
```

Abstract to analyze:
```

```
{text}

**Your output (JSON only):**
    """).format(
      text=text
    )
```

Listing 3: Metadata extraction prompt for biorxiv dataset

B Experiment details

B.1 Datasets

We adopt two challenging, real-world datasets for our studies.

biorxiv (Hou et al., 2025; Hu et al., 2025) is a dataset of abstracts on the biorxiv preprint server. We take the processed datasets from Hu et al. (2025). The training set size is $n = 28,846$, and the 95% quntile of the context token length is 512.

OpenReview (Xie et al., 2024) is a dataset of paper reviews from ICLR2023, where we take the processed datasets from Xie et al. (2024) directly. The training set size is $n = 8396$, and the downstream prediction labels are review area (12 classes) and recommendation (5 classes).

C Additional experiments

C.1 IMDB initialization results

We additionally evaluate whether MAPLE provides a stronger initialization on IMDB, a standard benchmark used in prior DP synthetic text work. This study focuses on the initialization stage only, since our goal is to test the same failure mode studied in the main paper: whether unconditional `RANDOM_API` generations are poorly aligned with the target data distribution before private evolution begins.

Setup. We use Qwen2.5-7B-Instruct for both MAPLE and AugPE initialization. For MAPLE, we set the total privacy budget to $\varepsilon = 3$ and run AIM on IMDB metadata using the allocated metadata budget. The metadata schema contains sentiment, writing style, primary focus, film genre, reviewer expertise, and review length. We then generate a balanced synthetic set of 2,000 reviews, with 1,000 positive and 1,000 negative examples, using DP synthetic metadata and in-context examples retrieved from a pool of 50 donated examples. For AugPE, we use the standard initialization prompt, "Write a movie review with a positive/negative sentiment," with no metadata or in-context examples; no privacy budget is consumed because this comparison stops before PE voting.

Evaluation. We compare synthetic reviews against a balanced IMDB test reference set of 2,000 reviews using MAUVE. Following the AugPE-style evaluation setup, we use `stsb-roberta-base-v2` features with scaling factor 5 and 500 buckets.

Table 1: Initialization-only IMDB comparison. Higher MAUVE is better.

Method	Privacy budget used	MAUVE (%)
AugPE initialization	none	1.08
MAPLE initialization	total $\varepsilon = 3$	10.43

MAPLE improves the IMDB initialization MAUVE from 1.08% to 10.43%. We also observed that AugPE initialization tends to produce reviews with a concentrated token-length distribution that is far from the real IMDB distribution. MAPLE can incorporate review length as DP metadata, producing a more diverse initial population. These results provide additional evidence that metadata-guided initialization can substantially reduce the initial distribution mismatch before PE refinement.