Training Private and Efficient Language Models with Synthetic Data from LLMs

Da Yu[1] **Arturs Backurs**[2] **Sivakanth Gopi**[2] **Huseyin Inan**[2] **Janardhan Kulkarni**[2]
Zinan Lin [2] **Chulin Xie** [3] **Huishuai Zhang** [2] **Wanrong Zhang**[4]
[1] Sun Yat-sen University [2] Microsoft Research
[3] University of Illinois at Urbana-Champaign [4] Harvard University
yuda3@mail2.sysu.edu.cn
{arturs.backurs,sigopi,huseyin.inan,jakul,zinanlin,huzhang}@microsoft.com
chulinx2@illinois.edu
wanrongzhang@fas.harvard.edu

Abstract

Language models are pivotal in modern text-based applications, offering many productivity features like next-word prediction, smart composition, and summarization. In many applications, these models must be lightweight to meet inference time and computational cost requirements. Furthermore, due to the inherent sensitivity of their training data, it is essential to train those models in a privacy-preserving manner. While it is well established that training large models with differential privacy (DP) leads to favorable utility-vs-privacy trade offs, training lightweight models with DP remains an open challenge.

This paper explores the use of synthetic data generated from a DP fine-tuned large language model (LLM) to train lightweight models. The key insight behind our framework is that LLMs are better suited for private fine-tuning, and hence using the synthetic data is one way to transfer such capability to smaller models. Our framework can also be interpreted as doing *sampling based* Knowledge Distillation in DP setting. It's noteworthy that smaller models can be trained on synthetic data using non-private optimizers, thanks to the post-processing property of DP. We empirically demonstrate that our new approach significantly improves downstream performance compared to directly train lightweight models on real data with DP. For instance, using a model with just 4.4 million parameters, we achieve 97% relative performance compared to the non-private counterparts in both medical and conversational corpus.

1 Introduction

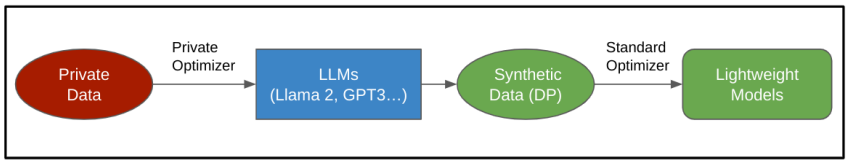


Figure 1: Illustration of the framework. We first fine-tune a large foundation model on private data with private optimizers. Then we use the trained LLM to generate DP synthetic data and train small models on the synthetic data using standard optimizers.

In modern text editors and email applications, lightweight language models play a crucial role in enhancing user experience through features such as next-word prediction and smart compose

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[Microsoft, 2020, Xu et al., 2023]. The data used to train these models is inherently sensitive, given the nature of the applications. Alarmingly, recent research has shown that deploying trained models may inadvertently leak information about their training data [Zhang et al., 2021, Carlini et al., 2021, Zhang et al., 2021, Matsumoto et al., 2023]. One promising way to mitigate these privacy risks is training the models with provable differential privacy [Abadi et al., 2016, Tramèr and Boneh, 2021, Bu et al., 2022b, De et al., 2022, Golatkar et al., 2022, Sander et al., 2022, He et al., 2023].

In recent years, there has been significant advancement in the training of large (and pre-trained) models with differential privacy [Anil et al., 2022, He et al., 2023, De et al., 2022, Golatkar et al., 2022, Bu et al., 2022a]. For instance, in the domain of natural language processing, Yu et al. [2022], Li et al. [2022b] demonstrate that a privately fine-tuned RoBERTa-Large model (355M) maintains over 96% relative accuracy compared to the non-private baseline on the GLUE benchmark [Wang et al., 2018]. Likewise, in computer vision, Berrada et al. [2023] exhibit that a privately fine-tuned NFNet-F7+ model (947M) retains more than 97% relative accuracy when compared to the state-of-the-art non-private results on ImageNet-1k [Deng et al., 2009]. Recently, He et al. [2023] showed that these findings hold true even for complex tasks such as summurization.

Although large foundation models can minimize the utility cost of DP, many real-world scenarios necessitate the use of efficient models, primarily to meet requirements related to inference speed and computational resources. For example, a language model supporting next-word prediction in a text editor must respond within a time limit, as otherwise users will type through by themselves. Furthermore, the adoption of smaller models also reduces the financial cost of deploying language models. However, despite these real-world demands, training efficient and private language models remains an open problem. Recent works indicate that achieving over 90% relative performance compared to the non-private counterpart is a major challenge for smaller models (typically with fewer than 50M parameters) [Kairouz et al., 2021, Wang et al., 2023].

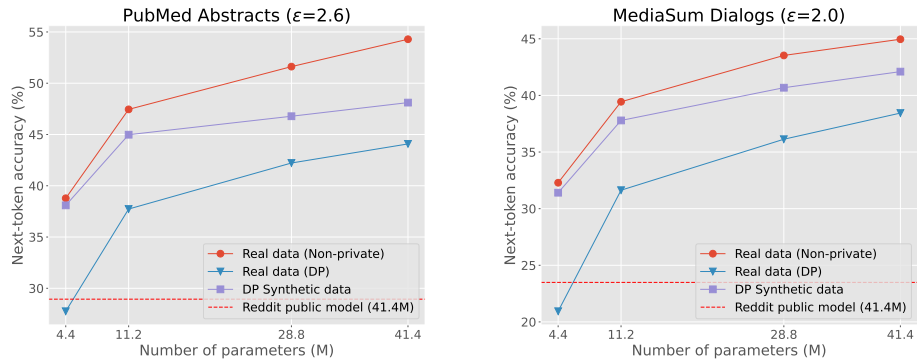
This paper explores the limits of training small models with DP. Our contributions are outlined below.

- We investigate a framework that uses synthetic text generated by privately fine-tuned large language models to train efficient and private downstream models. Figure 1 illustrates our framework. We expect that privately fine-tuned LLMs are powerful enough, allowing lightweight downstream models to perform well when trained on their synthetic data. In Appendix B, we discuss the insights behind our framework, which encompass recent findings in training large models with DP as well as the connection between our framework and knowledge distillation.
- Our experiments confirm this expectation. We evaluate our approach on two distinct types of datasets, abstracts of medical papers (targets at academic writing assistance) and conversation transcripts (targets at dialogue typing assistance). We found that models with just 4.4 million parameters, which can even be on-device models, retain more 97% relative performance compared to non-private baselines in both datasets. The main findings of our experiments are depicted in Figure 2. In Figure 5 and 6 in Appendix D, we present random samples of the synthetic datasets.
- Differentially private fine-tuning of Language Model Models (LLMs) can be a challenging task in terms of code implementation [He et al., 2023]. To the best of our knowledge, there is currently no opensource implementation for DP fine-tuning that supports model parallelism, which is crucial for training models with billions of parameters. In our code, we implement Fully Sharded Data Parallel [Zhao et al., 2023], enabling private fine-tuning of a 7B model with only four V100 GPUs. We plan to open source our implementation.

The rest of this paper is organized as follows. We discuss related work in Section 1.1 and preliminaries in Section 2. Implementation details as well as main results are depicted in Section 3. We conduct ablation study on some design choices in Section 4 and conclude in Section 5.

1.1 Related Work

Recent studies have explored the use of (large) language models for generating differentially private synthetic text [Bommasani et al., 2019, Putta et al., 2022, Mattern et al., 2022]. In recent studies by Yue et al. [2022], Kurakin et al. [2023], which focus on the sentiment classification task, it is shown that fine-tuning BERT/GPT-2 models on DP synthetic text produces similar performance



(a) PubMed Abstracts (b) MediaSum

Figure 2: Comparison of the two methodologies illustrated in Figure 1 in terms of next-token prediction accuracy of downstream models. Across a range of model sizes investigated, using DP synthetic text generated from a fine-tuned LLM consistently demonstrates superior performance.

to fine-tuning them on real data, underscoring the high-quality of the synthetic data. The main conceptual difference between these works and ours is that we use synthetic data as a framework to train smaller models and *a technique for DP model compression*, whereas previous works focused on generating synthetic data as their end goal. Further, compared to previous studies, our experiments introduce two novel contributions. We focus on text generation tasks, i.e., training the downstream model with next-token prediction loss, which is arguably more challenging than text classification. More importantly, we focus on lightweight downstream models, with the smallest model having only 4.4 million parameters. In this context, the use of DP synthetic text not only matches the performance of using real data but also yields significantly better results.

Recent research has also explored how to make DP models more efficient [Mireshghallah et al., 2022, Yu et al., 2023]. Mireshghallah et al. [2022] investigate private adaptations of knowledge distillation and model pruning techniques to compress a privately fine-tuned language model. Yu et al. [2023] demonstrate that the careful selection of pre-training data can enhance the private training of lightweight language models. Our study is different from this line of research in two key aspects. Conceptually, our main departure lies in demonstrating the feasibility of utilizing large foundation models to aid in the private training of lightweight models. Experimentally, for models with only 4.4 million parameters that are suitable for on-device applications, we are the first to establish that it is possible to maintain over 97% relative accuracy compared to the non-private counterpart.

2 Preliminaries

This section introduces the formal definition of differential privacy and some basics in differentially private deep learning.

Definition 1 ((ϵ, δ)-DP [Dwork et al., 2006, 2014])**.** *A randomized algorithm $\mathcal{M}$ is (ϵ,δ)-differentially private if for any two neighboring datasets D, D' and for every subset $\mathcal{S}$ of possible outputs:*

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

Two datasets D, D' are neighboring datasets if they differ in one sample. Specifically, following previous work, we assume that D can be transformed into D' by adding/removing exactly one sample.

DPSGD Algorithm. To implement differentially private deep learning, we use the DPSGD algorithm in Abadi et al. [2016]. It makes the gradients of SGD/Adam optimizers differentially private. At each step, the algorithm in Abadi et al. [2016] first clips per-example gradients to control the maximum contribution from each datapoint and then adds Gaussian noise to ensure that the contribution of a single example is concealed. For privacy accounting, we use the Privacy Random Variable (PRV) Accountant which gives a tighter privacy bound through numerical composition [Gopi et al., 2021, Ghazi et al., 2022, Koskela et al., 2020].

3 Private Training of Efficient Language Models Made Easy

We first introduce some basic experimental setup in Section 3.1 Other implementation details are documented in Appendix A. Then we present the main results in Section 3.2.

3.1 Setup

Models. We use the 7B version of Llama 2 [Touvron et al., 2023] as the LLM for generating synthetic text. The downstream models are four small Transformer models released by Turc et al. [2019]. The model sizes are 4.4M, 11.2M, 28.8M, and 41.4M (including the embedding matrix). The downstream models use WordPiece tokenizers and are pre-trained on Wikipedia and BookCorpus with masked language modeling. During fine-tuning, we apply a causal language modeling mask so that each token can only attend to its preceding tokens.

Datasets. We train next-token prediction models on PubMed abstracts and MediaSum, targeting at next-word completion in academic writing and daily conversation. Details are as follows.

PubMed abstracts. This dataset consists of the abstracts of medical papers from the National Library of Medicine[1]. We crawl the abstracts that were published between 2023/08/01 and 2023/08/07. The dates are after the cutoff date of the training data of Llama 2. The training set consists of abstracts that were published from 08/01 to 08/05. The validation set and the test set consist of abstracts that were published on 08/06 and 08/07, respectively. The training, validation, and test datasets have 75329, 4453, and 14423 abstracts, respectively. We generate 750K synethtic abstracts for training the downstream models.

MediaSum. MediaSum [Zhu et al., 2021] contains 463.6K transcripts collected from interview transcripts and overview/topic descriptions from National Public Radio and Cable News Network. Each sample of MediaSum consists of a dialog and an abstractive summary. We take the dialogs as our training data. To reduce computational cost, we only take dialogs that are shorter than 1K words. The training and test datasets have 182034 and 8245 dialogs, respectively. We generate 500K synthetic dialogs for training the downstream models.

Table 1: Next-token prediction accuracy (%) of downstream models. Privacy parameters are $(2.6, 10^{-5})$-DP for PubMed abstracts and $(2, 10^{-7})$-DP for MediaSum. The 'Δ' rows depict the improvements over training the models on real data with private optimizers. The abbreviation 'N.P.' stands for non-private. Subscripts numbers are the standard deviation calculated over three runs.

Parameters	4.4M	11.2M	28.8M	41.4M
PubMed Abstracts				
Real Data (N.P.)	$38.78_{0.038}$	$47.45_{0.027}$	$51.62_{0.014}$	$54.29_{0.033}$
Synthetic Data (N.P.)	$38.20_{0.068}$	$45.18_{0.072}$	$47.06_{0.083}$	$48.31_{0.073}$
Real Data	$27.76_{0.067}$	$37.73_{0.088}$	$42.22_{0.156}$	$44.08_{0.091}$
Synthetic Data	$38.09_{0.062}$	$44.98_{0.059}$	$46.78_{0.050}$	$48.11_{0.049}$
Δ	+10.33	+7.25	+4.56	+4.03
MediaSum Dialogs				
Real Data (N.P.)	$32.29_{0.016}$	$39.44_{0.014}$	$43.53_{0.013}$	$44.96_{0.020}$
Real Data	$20.94_{0.035}$	$31.63_{0.048}$	$36.18_{0.039}$	$38.44_{0.031}$
Synthetic Data	$31.41_{0.025}$	$37.79_{0.032}$	$40.68_{0.028}$	$42.10_{0.035}$
Δ	+10.47	+6.16	+4.50	+3.66

3.2 Main Results

We report the next-token prediction accuracy of downstream models in Figure 2 and Table 1. We evaluate two different methods for the private training of downstream models. The first one directly trains the downstream models on private data with private optimizers. The second one first trains Llama 2 7B on private data with private optimizers. Then it uses the fine-tuned Llama 2 7B to generate differentially private synthetic data. Finally, it trains the downstream models on the synthetic

[1] `https://www.ncbi.nlm.nih.gov/`

data with standard optimizers. In addition to fine-tuning the downstream models on PubMed abstracts or MediaSum, we also evaluate fine-tuning the models on an out-of-domain dataset that consists of 150K Reddit posts. The 150K Reddit posts are a subset of the Webis-TLDR-17 [Völske et al., 2017].

As shown in Figure 2, training the downstream models on differentially private synthetic data clearly outperforms other alternatives. When the model size is 4.4M, which is a typical size for on-device models, the proposed method achieves more than 98% relative accuracy of the non-private baseline on PubMed abstracts with $\varepsilon = 2.6$ and more than 97% relative accuracy accuracy of the non-private baseline on MediaSum Dialogs with $\varepsilon = 2.0$. Another observation is that the models fine-tuned on Reddit posts are worse than models fine-tuned on in-domain data. For example, 11.2M models fine-tuned on in-domain data outperform 41.4M Reddit models by a large margin.

Although the proposed method achieves better performance for all the downstream model sizes we evaluated, the results in Table 1 suggest that the improvement becomes smaller as the model size increases. For instance, on MediaSum dialogs, the absolute improvement of using DP synethtic text is 10.47% for the 4.4M downstream model while is only 3.66% for the 41.4M model. We anticipate that as the downstream model size keeps increasing, training the downstream model on real data directly will match the performance of training them on synthetic data.

4 Ablation Study

Here we run experiments on the PubMed abstracts dataset to investigate the impact of the number of synthetic samples, i.e., how many abstracts are generated using the fine-tuned LLM. We also run experiments with using different choices of top-k during inference to study how the diversity of the generated text affects downstream performance. A large top-k would increase the generation diversity. Due to space limit, the results are put in Appendix C.

Varying the number of synthetic samples. Our synthetic dataset for PubMed abstracts consists of three 250K subsets that are generated with different top-k values. To get synthetic datasets of different sizes, we take random subsets from the merged 750K dataset. We take three subsets with sizes 50K, 150K, and 400K. We present the results of training the downstream models on two types of synthetic data: non-private synthetic data, generated from a LLM fine-tuned without DP, and private synthetic data. Figure 3 shows the results of two downstream model sizes (4.4M and 41.4M). The performance of both models increases with the number of synthetic abstracts. Moreover, the larger model benefits more from a larger training set. Increasing the size of the synthetic dataset from 50K to 750K improves the next-token accuracy of the 41.4M model from 42.27% to 48.15%.

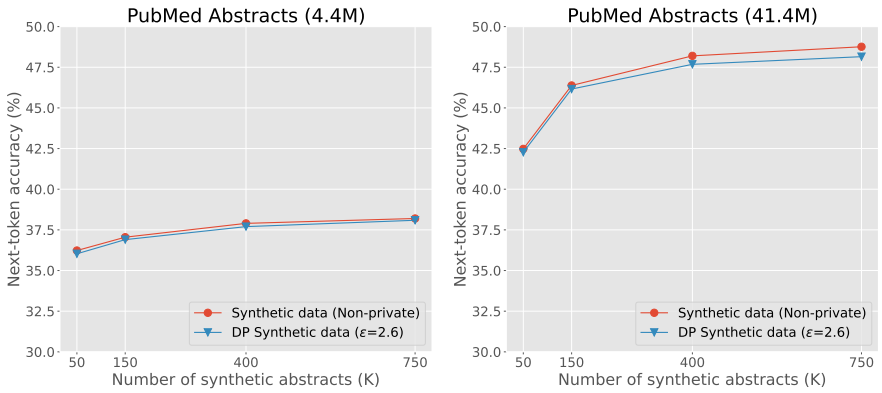


Figure 3: Accuracy of downstream models trained on four different sizes of synthetic datasets.

5 Conclusion

We explore utilizing DP synthetic text generated by large foundation models to enhance the private training of lightweight LMs. Notably, we demonstrate that small models with just 4.4 million parameters can maintain over 97% relative accuracy compared to their non-private counterparts. Our

computational resources restrict us to using foundation models up to 7 billion parameters. We expect that using more powerful foundation models could further strengthen the findings.

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*, pages 308–318, 2016.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private bert. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 2019.

Leonard Berrada, Soham De, Judy Hanwen Shen, Jamie Hayes, Robert Stanforth, David Stutz, Pushmeet Kohli, Samuel L Smith, and Borja Balle. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

Rishi Bommasani, Steven Wu, and Xanda Schofield. Towards private synthetic text generation. In *NeurIPS 2019 Machine Learning with Guarantees Workshop*, 2019.

Zhiqi Bu, Jialin Mao, and Shiyun Xu. Scalable and efficient training of large convolutional neural networks with differential privacy. *Advances in Neural Information Processing Systems*, 2022a.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term only fine-tuning of foundation models. *arXiv preprint arXiv:2210.00036*, 2022b.

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. *USENIX Security Symposium*, 2021.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 248–255, Washington, DC, USA, 2009. IEEE Computer Society.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 2014.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? *arXiv preprint arXiv:2302.09483*, 2023.

Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Faster privacy accounting via evolving discretization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022*, 2022.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 2021.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. *International Conference on Learning Representations*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pages 5213–5225. PMLR, 2021.

Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In Silvia Chiappa and Roberto Calandra, editors, *International Conference on Artificial Intelligence and Statistics*, 2020.

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.

Xuechen Li, Daogao Liu, Tatsunori B Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin-Tat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? *Advances in Neural Information Processing Systems*, 2022a.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *International Conference on Learning Representations*, 2022b.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023.

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

Microsoft. Assistive ai makes replying easier, 2020. URL `https://www.microsoft.com/en-us/research/group/msai/articles/assistive-ai-makes-replying-easier-2/`.

Fatemehsadat Mireshghallah, Arturs Backurs, Huseyin A Inan, Lukas Wutschitz, and Janardhan Kulkarni. Differentially private model compression. *Advances in Neural Information Processing Systems*, 2022.

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

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd. *arXiv preprint arXiv:2210.03403*, 2022.

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.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). *International Conference on Learning Representations*, 2021.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

Boxin Wang, Yibo Jacky Zhang, Yuan Cao, Bo Li, H Brendan McMahan, Sewoong Oh, Zheng Xu, and Manzil Zaheer. Can public large language models help private cross-device federated learning? *arXiv preprint arXiv:2305.12132*, 2023.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher A Choquette-Choo, Peter Kairouz, H Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. *arXiv preprint arXiv:2305.18465*, 2023.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Zhang Huishuai. Differentially private fine-tuning of language models. *International Conference on Learning Representations*, 2022.

Da Yu, Sivakanth Gopi, Janardhan Kulkarni, Zinan Lin, Saurabh Naik, Tomasz Lukasz Religa, Jian Yin, and Huishuai Zhang. Selective pre-training for private fine-tuning. *arXiv preprint arXiv:2305.13865*, 2023.

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

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*, 2021.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*, 2021.

Table 2: Hyperparameters for fine-tuning Llama 2 7B and downstream models. For fine-tuning Llama 2, we only tune the hyperparameters on PubMed abstracts and reuse the best configuration on MediaSum dialogs. For a target ε, noise multiplier is set as the smallest value such that DP-SGD can run target number of steps.

	Llama 2	Downstream (pri.)	Downstream (non-pri.)
Epoch	5	[10, 30, 50, 100]	[10, 30]
Batchsize	[512, 1024, 2048]	[512, 1024, 2048]	64
Clipping norm	[0.1, 1]	[0.1, 1, 5]	N/A
Learning rate		[3e-5, 1e-4, 3e-4, 1e-3, 3e-3]	

A Hyperparameters

Here we document all the hyperparameters in our experiments. Our source code will be publicly available after the review process.

Llama 2 7B. For fine-tuning Llama 2, we use LoRA, a parameter-efficient fine-tuning algorithm [Hu et al., 2022, Yu et al., 2022], to reduce the computational cost. We apply LoRA with a rank of 16. This gives 17.5M trainable parameters. The hyperparameters for fine-tuning Llama 2 on PubMed abstracts are in Table 2. For fine-tuning Llama 2 on MediaSum dialogs, we directly transfer the best hyperparameters we found on PubMed abstracts. The learning rate is 1e-3, the clipping norm is 1, the noise multiplier is 1.1, the batch size is 2048, and the number of epochs is 10. This ends up with the ε equal to 2.0.

We use unconditional generation to generate synthetic abstracts and dialogs. The number of synthetic MediaSum dialogs is 500K, generated with top-$k = 200$. The top-k parameter controls the diversity of the generated text. It determines the size of the candidate pool from which the model selects the next token and a large value of top-k would increase the generation diversity. The number of synthetic synthetic PubMed abstracts is 750K. It consists of three subsets are generated with different top-k (50, 200, and ∞). We use different choices of top-k to study the impact of top-k sampling on downstream accuracy.

Downstream models. The hyperparameters for fine-tuning all downstream models (4.4M, 11.2M, 28.8M, and 41.4M) are also in Table 2. When fine-tuning with differential privacy, the privacy parameters are $\varepsilon = 2.6$ and $\delta = 1 \times 10^{-5}$ for PubMed abstracts and $\varepsilon = 2.0$ and $\delta = 1 \times 10^{-6}$ for MediaSum dialogs. All downstream experiments are repeated 3 times with different random seeds.

B Insights Behind Our Framework

Here we give a brief theoretical interpretation on why our framework should be broadly applicable for training smaller models across a range of parameters and applications. We emphasize that this interpretation is not rigorous, but what guided us in understanding our framework. There are two ways to look at our framework.

1. **Larger pre-trained models are better suited for private fine-tuning.** Recent research indicates that employing differentially private training methods on large models results in a smaller cost in utility. This phenomenon has been substantiated both empirically [Li et al., 2022b, Yu et al., 2022, He et al., 2023, Bu et al., 2022b] and theoretically [Li et al., 2022a, Ganesh et al., 2023]. Consequently, leveraging synthetic data generated through a privately fine-tuned large model is a viable approach to transfer such capabilities to smaller models.
2. **Knowledge Distillation via Sampling:** The main idea behind KD algorithm Hinton et al. [2015] is to train a student model to produce a distribution P_s to mimic the output distribution of the teacher P_t for a given input. The intuition is that output distribution of the teacher captures the knowledge learnt by the teacher, in particular probabilities assigned by the teacher to tokens that are different from the true next token. In particular, one trains to minimize the cross entropy loss between the distribution of the student and the student: -$\sum_{y \in Y} P_t(y) \cdot \log(P_s(y))$. A sampling based interpretation of this above objective West et al. [2021] can be thought of as minimizing $E_{y \sim P_t(y)}[-\log(P_s(y))]$. Intuition tells us

that as we generate more samples from the teacher, the latter objective, in the limit, should approximate the more standard KD objective.

KD algorithm is one of the most widely used algorithms for transferring knowledge from a bigger network to a smaller network. However, Mireshghallah et al. [2022] showed that it is difficult to adapt KD algorithm to the DP setting. Our framework can be thought of as doing KD algorithm on the samples from the teacher, and our experiments indicate that our framework can be an effective way of invoking KD in the DP setting. Our experiments also validate that as the synthetic dataset size increases, performance of the student improves, as should be clear from the discussion above.

C Varying the Diversity of the Generated Text.

Here we study the impact of the diversity of synthetic text on the downstream performance. We change the top-k parameter to change the generation diversity. The top-k parameter decides the candidate pool used for sampling the next token, with a larger top-k value leading to heightened diversity. We use three different values of top-k, 50, 200, and ∞. The next-token accuracy of four different downstream models are in Figure 4. When the Llama 2 model is fine-tuned non-privately, all three choices of top-k yield similar downstream accuracy. However, when the model is fine-tuned with DP, using top-$k = \infty$ leads to worse downstream accuracy. One plausible explanation for this observation is that training with (DP) has a more pronounced influence on the less frequent vocabulary items [Bagdasaryan et al., 2019, Feldman, 2020]. Consequently, sampling from the entire vocabulary may reduce the quality of synthetic text.

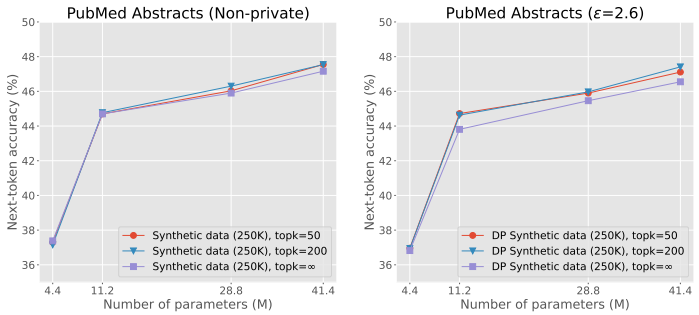


Figure 4: Impact of varying top-k choices on downstream accuracy. The left plot shows the results of using non-private synthetic data. The right plot shows the results of using private synthetic data.

D Random Samples in the Synethtic Datasets

Two Random Samples in the Synthetic Pubmed Dataset

Pulmonary edema in pulmonary arterial hypertension (PAH) is associated with increased mortality and morbidity. The development of edema in PAH may occur via direct vascular damage, by the release of pro-inflammatory cytokines, and by the changes associated with vasoconstriction. The role of systemic immunopharmacological drugs in disease course control has to be fully clarified. In the recent years, the development of biologics that can be used as monotherapy or in combinations has markedly influenced the management of PH. At the same time, the use of inhaled prostacyclins, calcium channel blockers, and inhaled nitrates has had a major impact on the development of PAH. In addition, non invasive imaging techniques have been developed to assist the assessment of the impact of therapy on cardiac and vascular hemodynamics. In this article, we discuss the current state of the treatment of PAH by inhaled prostacyclin analogues, ions channel blockers, and inhaled nitrates, the potential of biological therapies, the use of imaging techniques for assessment of disease course, and the persistence of residual questions and concerns about the therapies available. This is a comprehensive review of the treatment and mechanisms of disease progression in pulmonary arterial hypertension.

In contrast to prior work, which has generally focused on high-cost pharmaceuticals sold via private medical-surgical wholesalers, this study examines the impact of the wholesaler consolidation on the distribution of low-cost pharmaceuticals that are mainly sold via retail pharmacies and large-chain mail-order pharmacies. Specifically, utilizing the 2012 Pharmaceutical Distributors Sales Data (PDSD), we use econometric techniques to examine the wholesaler consolidation pattern among non-Specialty Distributors among the top 50 non-Specialty Distributors and across each region in 2011, 2016, 2017, and in 2018, respectively. Our results show that the wholesaler consolidation pattern increased the market power of the top distributors and resulted in an increase in prices, including an 11.3% increase in the price of commonly used and low-cost generic drugs. Furthermore, we find significant differences in market concentration and competition across different regions. The findings of this study have implications for the healthcare system and highlight the need for policies to ensure fair competition between distributors.

Figure 5: Two random samples from the synthetic PubMed abstracts. The generator is trained with ε =2.6 and $\delta = 1 \times 10^{-5}$. The synthetic samples are generated with top-k=50.

One Random Sample in the Synthetic Dialogs

```
JASON CARRO: We have a report. It's not an overly warm one, but they're still making snow in
the Swiss Alps.
JASON CARRO, CNN CORRESPONDENT: You have your skis, your snow shoes, and not a lot of snow in
the Swiss Alps near Jungfraujoch.
CARRO (voice-over): The Jungfrau region is famous for its snow-covered mountains and the
local ice hotels. However, the cold winter weather this season has forced some of the hotels
to close temporarily. In the months of December and January, heavy snowfall dropped only 14
inches. Despite the concerns over the lack of snow, the local ski resort reports business is
still good.
DANIEL TISO: Skiers and the tourists didn't care about snow, but it helped us to spread the
word. We are famous for no-snow and yes, it's very good.
CARRO: Jungfraujoch is usually reputed as the top of Europe, at the end of the mountain range,
but after heavy snow, it's shrinking.
CARRO (voice-over): The Ice Palace was one of the many attractions.
(voice-over): And at nearly 12,000 feet above sea level, you can get a view of Europe's ski
spots.
CARRO (on camera): It's true you don't always think Switzerland when you think of skiing, but
it's all about the snow there for them.
TISO: We have many things more than snow, but that is the main thing.
CARRO (voice-over): That's Daniel Tiso (ph). He's been skiing when he was a boy, now he's
taking his grandchildren, and he says he would not have it any other way, especially not now
that it's so warm without any snow.
TISO (ph): You can still ski over 100,000 people around the highest mountain in Europe, but
normally it's - it's completely empty.
CARRO (on camera): It's a trade-off, it's all about the snow - Jason Carroll, CNN.
```

Figure 6: One random sample from the synthetic dialogs. The generator is trained with ε =2.0 and $\delta = 1 \times 10^{-6}$. The synthetic samples are generated with top-k=50.