Fine-Tuning Language Models with Differential Privacy through Adaptive Noise Allocation

Xianzhi Li[1], Ran Zmigrod[2], Zhiqiang Ma[2], Xiaomo Liu[2], Xiaodan Zhu[1]
[1]Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute
Queen's University
[2]J.P. Morgan AI Research
{li.xianzhi, xiaodan.zhu}@queensu.ca
{ran.zmigrod, zhiqiang.ma, xiaomo.liu}@jpmchase.com

Abstract

Language models are capable of memorizing detailed patterns and information, leading to a double-edged effect: they achieve impressive modeling performance on downstream tasks with the stored knowledge but also raise significant privacy concerns. Traditional differential privacy based training approaches offer robust safeguards by employing a uniform noise distribution across all parameters. However, this overlooks the distinct sensitivities and contributions of individual parameters in privacy protection and often results in suboptimal models. To address these limitations, we propose ANADP, a novel algorithm that adaptively allocates additive noise based on the importance of model parameters. We demonstrate that ANADP narrows the performance gap between regular fine-tuning and traditional DP-SGD based fine-tuning on a series of datasets while maintaining the required privacy constraints.

1 Introduction

Language models have achieved remarkable success and shown impressive abilities in a wide range of tasks (Almazrouei et al., 2023; Touvron et al., 2023; Team et al., 2023). Their advanced capabilities in memorizing detailed information and patterns in data as well as making connections among them have not only helped language models to achieve impressive modeling performance on downstream tasks, but also raised significant and ubiquitous privacy concerns if it is not properly handled (Neel and Chang, 2023; Mireshghallah et al., 2023; Yao et al., 2024).

Differential Privacy (DP) is a principled framework for mitigating privacy risks, providing theoretical guarantees that prevent inferring the presence or absence of an individual's data in a model's output (Abadi et al., 2016; Dwork, 2006). Conventional DP-enhanced fine-tuning offers robust safeguards by assuming a uniform noise distribution across all parameters to protect privacy (Kerrigan et al., 2020; Yu et al., 2021b; Li et al., 2021). Recent work has begun exploring DP in Parameter Efficient Fine Tuning (PEFT) (Yu et al., 2021a; Bu et al., 2022), which is built on the same assumption on the additional tunable parameters. Unfortunately, such an assumption overlooks the distinct sensitivities and contributions of individual parameters in privacy protection and often results in suboptimal models.

In this paper, we introduce ANADP, a novel DP method that adaptively distributes the noise and privacy budget among a language model's parameters during fine-tuning, based on their importance to the model at a given training step. Our work was inspired by Zhang et al. (2023), who utilized the sensitivity and uncertainty of parameters for model pruning. Importantly, our approach not only respects the inherent heterogeneity of parameter significance but also maintains strong privacy protection. The proposed integration addresses the key challenges in effectively measuring the contributions of parameters and ensures that models are trained stably. We demonstrate that ANADP consistently improves performance over the traditional DP fine-tuning under the same privacy budget and bridges the gap between traditional DP and non-DP fine-tuning (no privacy guarantee). The contributions of our work are summarized below:

- We propose ANADP, a novel algorithm for fine-tuning language models while maintaining privacy guarantees. To the best of our knowledge, this is the first DP method that distributes the privacy budget based on Transformer parameters' importance non-uniformly.
- We empirically demonstrate that ANADP outperforms the standard DP approaches on the Glue benchmark (Wang et al., 2018) in multiple training paradigms (e.g. both full fine-tuning and PEFT).

Findings of the Association for Computational Linguistics: EACL 2024, pages 8368–8375

- We conduct further analysis on privacy exposure risk and find that ANADP offers the same robust privacy protection as the conventional DP method.

2 Related Work

Differential Privacy. DP is a principled approach to ensuring privacy. The concept of DP was formalized by Dwork (2006), who introduced the definition and foundational mechanisms of DP. In machine learning, Abadi et al. (2016) introduced the widely used Differentially Private Stochastic Gradient Descent (DP-SGD). Adaptive Differential Privacy is a recent development. Research (Gong et al., 2020; Chen et al., 2023) have been developed to preserve adaptive DP in deep neural networks. However, these methods fall short in capturing the complex parameter interactions within transformers, potentially leading to suboptimal models and trade-offs between privacy and utility.

Fine-Tuning and PEFT. Full fine-tuning used to be a prominent approach but can be resource-intensive and less efficient (Lester et al., 2021; Tay et al., 2022). PEFT has emerged as another option for effectively training LLMs. Many PEFT techniques, such as LoRA (Hu et al., 2021), Adapters (Houlsby et al., 2019), and prefix tuning (Li and Liang, 2021) have been proposed to tune small, additional modules instead of the whole model. He et al. (2021) provided a unified view revealing the connections among various parameter-efficient transfer learning methods. Recent work by (Zhang et al., 2023) introduced AdaLoRA to dynamically adjust the amount of parameter tuning based on the task and model requirements. Despite their efficiency, integrating these methods with privacy-preserving techniques remains an area that requires further exploration.

3 The ANADP Model

The integration of Differential Privacy for language model fine-tuning is crucial for deploying LLMs in privacy-sensitive applications. In this work, we introduce ANADP, an adaptive noise allocation DP training method based on the importance score of models' parameters, which provides a generic solution that can be applied to a wide range of LLMs. The fundamental idea is that adding less noise to the parameters that are more important and more to the less important parameters would help improve the model's utility given the same privacy budget. This section describes the construction and correctness of ANADP, whose pseudocode is given in Alg. 1.

Algorithm 1 ANADP Algorithm

Input: Training batches $\mathcal{L} = \{L_1, ..., L_T\}$, Initial parameter weighs ω_0, noise multiplier σ_0
Hyper-parameters: α, β_1, β_2, clipping threshold C, learning rate γ
$S_0 \leftarrow \mathbf{0}$, $\overline{S}_0 \leftarrow \mathbf{0}$, $\overline{U}_0 \leftarrow \mathbf{0}$
for $L_t \in \mathcal{L}$ **do**
  Compute gradients $g(L_t)$
  $S_t \leftarrow |g(L_t) \cdot \omega_{t-1}|$ ▷ *Compute Sensitivity*
  $\overline{S}_t \leftarrow \beta_1 \overline{S}_{t-1} + (1-\beta_1) S_t$ ▷ *Eq. 3*
  $\overline{U}_t \leftarrow \beta_2 \overline{U}_{t-1} + (1-\beta_2)\left|\overline{S}_t - S_t\right|$ ▷ *Eq. 4*
  $I_t \leftarrow \overline{S}_t \cdot \overline{U}_t$ ▷ *Eq. 5*
  $\mu \leftarrow \text{mean}\left(\frac{I_t - \text{median}(I_t)}{q_1(I_t) - q_2(I_t)}\right)$ ▷ *Mean importance*
  $\widehat{I}_t \leftarrow (1-\alpha)\left(\frac{I_t - \text{median}(I_t)}{q_1(I_t) - q_2(I_t)}\right) + \alpha\mu$ ▷ *Eq. 6*
  $\overline{I}_t \leftarrow \widehat{I}_t - \left(\text{mean}\left(\widehat{I}_t\right) - 1\right)$ ▷ *Eq. 7*
  $\tilde{g}(L) \leftarrow \min\left(g(L), C\right) + \mathcal{N}\left(\frac{\sigma_0^2}{\overline{I}_t}\right)$ ▷ *Eq. 8*
  $\omega_t \leftarrow \omega_{t-1} - \gamma g(L_t)$ ▷ *Update weights*
end for
Output: Updated parameters ω_T

We follow Dwork (2006)'s definition of DP. Specifically, we achieve DP through a **randomized algorithm** A over an output space $\mathcal{S}$. Given a privacy budget ϵ and error probability δ, we say A is **(ϵ, δ)-differentially private** (**(ϵ, δ)-DP**) if for any neighboring datasets D and D', which differ in exactly one data record, the following inequality holds:

$$\Pr\left[A(D) \in \mathcal{S}\right] \leq e^{\epsilon} \Pr\left[A(D') \in \mathcal{S}\right] + \delta \quad (1)$$

where privacy budget ϵ is a measure of the amount of privacy loss allowed during training. Past methods for achieving (ϵ, δ)-DP typically add a uniform Gaussian noise to the parameters. More formally, given a batch L, we can define adding Gaussian noise to the model's gradient, $g(L)$ as:

$$\tilde{g}(L) \stackrel{\text{def}}{=} \min\left(g(L), C\right) + \mathcal{N}\left(C\sigma^2\right) \quad (2)$$

where C is a clipping threshold and $\mathcal{N}\left(C\sigma^2\right)$ is Gaussian noise with mean 0 and variance $C\sigma^2$. C and σ are fixed and computed based on the privacy budget (Abadi et al., 2016). In this work, we explore a tunable σ to realize a better trade-off between privacy and utility. We aim to tailor the noise

distribution across different parameters and the key objective is to determine the importance of each parameter.

Parameter Importance. In order to gauge parameters' importance, our work is inspired by Zhang et al. (2023), which calculates importance based on the sensitivity and uncertainty of the parameter for model pruning. We use the moving averages of the sensitivity and uncertainty of the model parameters at training step t:

$$\overline{S}_t \stackrel{\text{def}}{=} \beta_1 \overline{S}_{t-1} + (1-\beta_1) S_t \tag{3}$$

$$\overline{U}_t \stackrel{\text{def}}{=} \beta_2 \overline{U}_{t-1} + (1-\beta_2) \left| \overline{S}_t - S_t \right| \tag{4}$$

where $\beta_1, \beta_2 \in [0,1]$ are hyper-parameters to control the move average rate. Additionally, $S_t \stackrel{\text{def}}{=} |g(L_t) \cdot \omega_{t-1}|$ is the **sensitivity** of the model weights at step t.[1] The **importance metric** is then the element-wise product of the sensitivity and uncertainty

$$I_t \stackrel{\text{def}}{=} \overline{S}_t \cdot \overline{U}_t. \tag{5}$$

This formulation ensures that parameters with moderate sensitivity but high uncertainty are still considered important, which prevents prematurely discarding parameters that could become important as training progresses.

Importance Normalization. Using I_t as defined in Eq. 5 may lead to zero-gradients. Therefore, we smooth the importance scores

$$\widehat{I}_t \stackrel{\text{def}}{=} (1-\alpha) \left(\frac{I_t - \text{median}(I_t)}{q_1(I_t) - q_2(I_t)} \right) + \alpha\mu \tag{6}$$

where $\alpha \in [0,1]$ is a smoothing parameter, $q_1(I_t)$ and $q_2(I_t)$ are chosen quantiles of I_t, and μ is the mean of the scaled normalized vector.

$\widehat{I}_t$ gives a scaled distribution of importance across the model's parameters. In order to ensure (ϵ, δ)-DP, we further adjust distribution to be centered at one

$$\overline{I}_t \stackrel{\text{def}}{=} \widehat{I}_t - \left(\text{mean}\left(\widehat{I}_t \right) - 1 \right). \tag{7}$$

This means that the overall noise added to the model will follow the same distribution as that of Abadi et al. (2016) who uses a uniform noise across all parameters. As the overall noise added is the same, ANADP satisfies the (ϵ, δ)-DP guarantees following the proof of Abadi et al. (2016). The smoothed importance score is utilized to adaptively add noise to the gradient. Replicating Eq. 2, our new gradient is:

$$\tilde{g}(L) = \min\left(g(L), C\right) + \mathcal{N}\left(\frac{\sigma_0^2}{\overline{I}_t}\right) \tag{8}$$

where σ_0 is a noise multiplier that achieves (ϵ, δ)-DP and is selected following Abadi et al. (2016).

4 Experiments

4.1 Experimental Setup

We evaluate the performance of ANADP against the traditional DP-SGD method (Abadi et al., 2016), DP-PEFT method (Yu et al., 2021a) as well as regular fine-tuning (i.e., no privacy guarantees). Same as in previous work (Wu et al., 2023; Yu et al., 2021a), we run our three privacy configurations using RoBERTa (base and large) (Liu et al., 2019) in the full fine-tuning setting as well as on two state-of-the-art PEFT methods: LoRA (Hu et al., 2021) and Adapter (Houlsby et al., 2019).[2] Each of the combinations above is evaluated against four datasets from the Glue benchmark (Wang et al., 2018) which is used in past work to evaluate conventional DP-SGD: SST-2 (Socher et al., 2013), QNLI (Rajpurkar et al., 2016), MNLI (Williams et al., 2018), and QQP.[3]

We further conduct privacy protection experiments. Our experiments follow those of (Wu et al., 2023); we train a model using contexts containing private information (the Enron email dataset (Klimt and Yang, 2004)), and then compute the leakage risk of the privacy information. The Enron email dataset is comprised of over 500,000 emails that contain sensitive information such as person names and phone numbers Specifically, we use the Mean Reciprocal Rank (MRR) for person name and exposure (Carlini et al., 2019) metric for telephone numbers to show the risk of privacy leakage.

4.2 Accuracy Results

The performance of ANADP in comparison to past DP methods and regular training is given in Table 1. Introducing privacy protection inevitably leads to performance degradation. Nevertheless, we observe that ANADP consistently outperforms

[1] The **uncertainty** of a parameter at step t is the absolute difference between its sensitivity at step t and its moving average $\overline{S}_t$.

[2] More details including hyperparameters chosen are given in Appendix A.

[3] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Method		Roberta-base					Roberta-large				
Paradigm	**Privacy Alg.**	**SST-2**	**QNLI**	**MNLI**	**QQP**	**Avg.**	**SST-2**	**QNLI**	**MNLI**	**QQP**	**Avg.**
Full	w/o DP	94.8	92.8	87.6	91.9	91.8	96.4	94.7	90.2	92.2	93.4
	DP	91.5	87.9	83.4	86.4	87.5	95.0	91.2	87.2	86.8	90.1
	ANADP	92.5	89.1	84.0	87.6	88.3	95.2	92.3	87.9	88.5	91.0
LoRA	w/o DP	95.1	93.3	87.5	90.8	91.7	96.2	94.9	90.6	91.6	93.3
	DP	92.2	87.3	83.5	85.7	87.2	95.3	90.8	87.8	87.4	90.3
	ANADP	93.4	88.8	83.9	86.5	88.2	95.7	91.9	88.1	87.7	91.0
Adapter	w/o DP	94.7	93.0	87.3	90.6	91.4	96.4	94.7	90.3	91.5	93.2
	DP	92.5	87.5	83.4	85.6	87.3	93.9	90.7	87.7	86.3	89.7
	ANADP	93.4	88.0	84.4	86.3	88.0	94.8	91.8	88.7	87.7	90.8

Table 1: Performance Comparison (accuracy) for ANADP with baselines using full fine-tuning, LoRA, and Adapter tuning. The performance differences between ANADP and DP are all statistically significant with $p < 0.05$ under the one-tailed paired t-test.

traditional DP-SGD full fine-tuning and PEFT fine-tuning, demonstrating the benefit of using ANADP. For instance, ANADP achieves a performance improvement of 1.4% for RoBERTa-large on the QQP task in the Adapter setting, and 1.5% for RoBERTa-base on the QNLI task in the LoRA setting. For full fine-tuning, ANADP also poses a performance gain of up to 1.7% compared to the conventional DP-SGD. The improvement is consistent across all the tasks and settings. In our additional experiments, we found that ANADP only introduces less than 1% more GPU memories and 5% more training time compared to the original DP method, yet achieves better utility. The performance differences between ANADP and DP are all statistically significant with $p < 0.05$ under one-tailed paired t-test. Finally, in order to better understand how ANADP differs from past DP techniques, we examine the detailed noise distribution breakdown in Figure 2.

4.3 Exposure Risks Results

Our exposure experiments seek to assess the risk of privacy leakage empirically, particularly focusing on sensitive information such as person names and telephone numbers. Such an experiment is important as concerns have previously been raised on whether DP guarantees adhere to the allocated privacy budget (Steinke et al., 2024). This discrepancy can occur due to various factors, ranging from theoretical assumptions not holding in practice to statistical variations and implementation bugs.

Figure 1 shows that ANADP maintains the same level of privacy protection as the conventional DP methods on the Enron email dataset (Klimt and Yang, 2004), without statistically significant difference between them. The exposure risk values show a substantial reduction compared to those

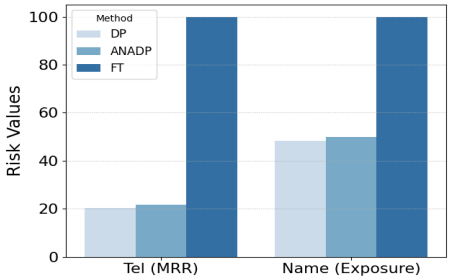

Figure 1: Privacy leakage risks for telephone number and person name using DP, ANADP and non-DP full fine-tuning.

of the non-DP model, demonstrating its effectiveness in mitigating privacy leakage risks. Overall, with the same privacy protection capability as conventional DP, ANADP consistently improves the performance of the latter in the benchmark tasks described above, benefiting from considering the different contributions of parameters.

4.4 ANADP Noise Distribution

Figure 2 shows the detailed distribution of noise multipliers applied via ANADP when tuning RoBERTa-base on the SST-2 task. ANADP demonstrates a strategic pattern in noise allocation, consistently applying lower noise levels to more critical parameters. Notably, the lower and final layers of the model often receive reduced noise. This could suggest that initial layers, responsible for capturing basic linguistic features, and final layers, which fine-tune these features into task-specific outputs, are deemed more sensitive to noise disruption. This pattern supports the hypothesis that maintaining the integrity of these parameters is crucial for preserving the model's overall performance.

In contrast, ANADP strategically assigns higher

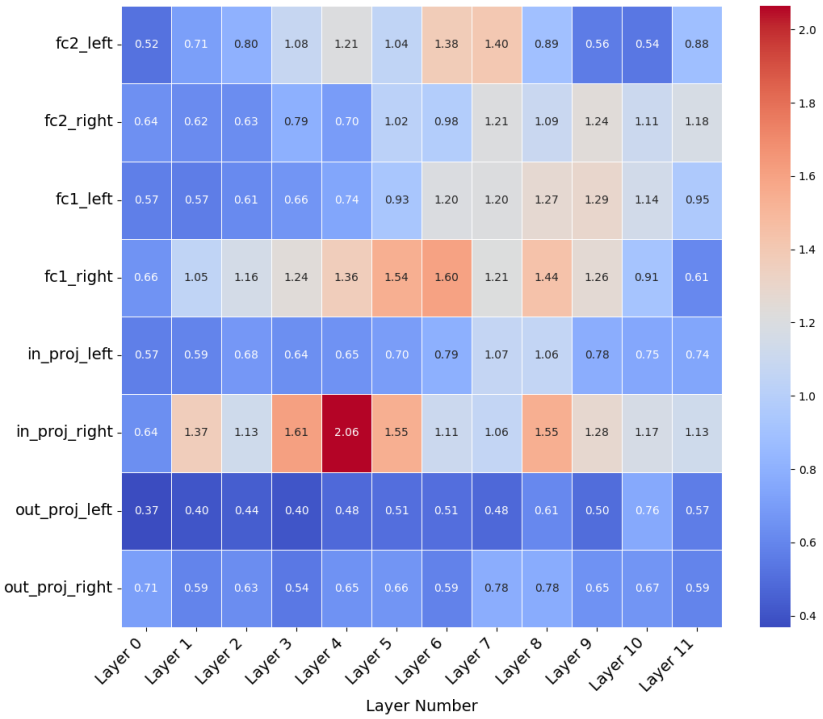

Figure 2: Distribution of noise multipliers during the training of ANADP on the SST-2 dataset. The X-axis represents the 12 layers of the Roberta-base model, while the Y-axis denotes the PEFT weights. The color gradient indicates the varying amounts of noise applied.

noise levels to the middle layers of the transformer model. This allocation aligns with the findings of Meng et al. (2023) which concluded that factual knowledge is predominantly stored in the middle layers of the feed-forward network. By introducing more noise to these layers, ANADP effectively obfuscates sensitive factual associations, thereby enhancing privacy protection without compromising the model's ability to learn and perform on specific tasks. This targeted noise allocation ensures that while the privacy of stored knowledge is robustly safeguarded, the overall performance of the model remains optimized.

5 Conclusion

This paper introduces ANADP, a novel approach to integrating DP with language model fine-tuning in both the full fine-tuning and PEFT settings and dynamically adjusting the noise added to the gradients, based on measuring model parameters' importance. We demonstrated that under the same privacy budget, ANADP consistently outperforms the standard DP-SGD training on different benchmark datasets. While performance degradation remains between our method and non-DP training, we achieved consistent reduction of the gap, in both the fine-tuning and PEFT settings. Our additional exposure risk analysis shows that ANADP provides privacy protection comparable to the standard DP-SGD training. We hope this work enables better deployment of privacy-preserving language models and encourages future research on adaptive DP for language model training.

6 Limitations

While ANADP offers consistent improvements, there are certain limitations that present opportunities for future work. First, although our method effectively identifies important parameters for downstream tasks and allocates noise accordingly, it does not explicitly distinguish whether these parameters are also privacy-sensitive. Identifying privacy-related parameters during the training process could be a crucial research problem. Moreover, developing an automated method to normalize noise would significantly streamline the application of ANADP.

Acknowledgement

This research was funded in part by the Faculty Research Awards of J.P. Morgan AI Research. The authors are solely responsible for the contents of the paper and the opinions expressed in this publication do not reflect those of the funding agencies.

Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

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

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Lin Chen, Danyang Yue, Xiaofeng Ding, Zuan Wang, Kim-Kwang Raymond Choo, and Hai Jin. 2023. Differentially private deep learning with dynamic privacy budget allocation and adaptive optimization. *IEEE Transactions on Information Forensics and Security*.

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

Maoguo Gong, Ke Pan, Yu Xie, A Kai Qin, and Zedong Tang. 2020. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition. *Neural Networks*, 125:131–141.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. *arXiv preprint arXiv:2009.05886*.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

Seth Neel and Peter Chang. 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2024. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *Preprint*, arXiv:2009.06732.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

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

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021b. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

A Appendix A

Training Details. Following prior work in model privacy, we conduct our training using RoBERTa (on both the *base* and *large* version) (Liu et al., 2019). RoBERTa-base has 12 transformer layers, a hidden state size of 768, and a feedforward network (FFN) with an internal hidden size of 3072. RoBERTa-large is configured with 24 transformer layers, enhancing its complexity. In LoRA fine-tuning, we followed Yu et al. (2021a) where we incorporated bottleneck branches in both the attention layers and the feedforward layers. This approach differs slightly from the method used by Hu et al. (2021), who only added bottleneck branches to the query and values matrices within the attention layers. For the two types of PEFT methods, we choose the same rank 16 for all the experiments. For DP experiments, we use ϵ=8, C=10, δ=1e-5 for SST-2, QNLI and δ=1e-6 for MNLI, QQP dataset. We run 50 epochs on all datasets and report the best validation accuracy. All experiments were conducted using NVIDIA A100 GPUs.

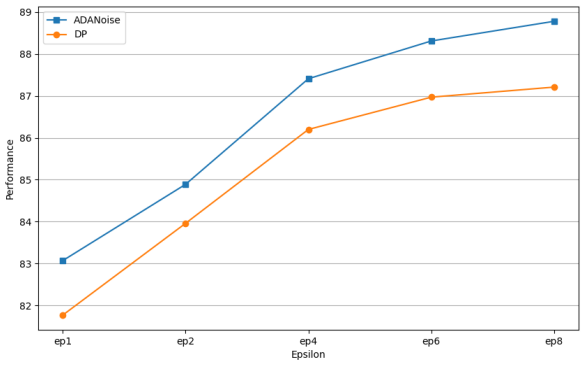

Figure 3: Performance of ANADP under different privacy budget on QNLI dataset.

We further extend our analysis by examining the performance of ANADP under different privacy budgets, as illustrated in figure 3. The trends demonstrate that ANADP consistently outperforms traditional DP under all scenarios, even when the privacy budget is set to a relatively low value ($\epsilon = 1$). This difference becomes more pronounced as ϵ increases, with ANADP reaching 88.78 at $\epsilon = 8$ compared to DP's 87.21. This consistent outperformance highlights the effectiveness of ANADP.

we have included additional experimental results from BERT in the table below. These results demonstrate the effectiveness and versatility of ANADP across various models, reinforcing its generalizability. It is also worth noting that, in the current studies on DP-based models, it is not common to use larger generative models, and most packages do not support such models. We follow the same setup to make our work comparable to the existing work. We leave the investigation on larger generative models as future work.

Table 2: Comparison of ANADPand DP with BERT-base.

Task	Method	BERT-base
SST-2	ANADP	88.18
	DP	87.27
QNLI	ANADP	86.61
	DP	86.03
MNLI	ANADP	79.46
	DP	78.85
QQP	ANADP	84.97
	DP	84.72