Is Privacy Noise Helpful Regularization or Harmful Optimization Noise in Small-Scale LLM Fine-Tuning?

Leitian Tao [1]

Abstract

Differentially private (DP) training protects the contribution of individual examples, but its effect on optimization is often difficult to interpret. In DP-SGD, per-example gradient clipping and additive Gaussian noise may have competing effects: clipping can stabilize updates and reduce overfitting, while noise can corrupt gradient estimates and slow convergence. This report studies this tension in small-scale language-model fine-tuning, a regime where privacy is practically important and optimization is fragile. We compare non-private fine-tuning, clipping-only fine-tuning, noise-only fine-tuning, and full DP-SGD across open-weight language models and varying training-set sizes. We use validation accuracy, train–validation gaps, loss trajectories, gradient-norm variability, and cross-seed variance to distinguish regularization-like behavior from harmful optimization noise. Our experimental results show that mild privacy noise can provide limited regularization benefits in the smallest-data regime, but stronger noise primarily degrades optimization.

1. Introduction

Differential privacy (DP) provides a formal way to limit how much information a trained model can reveal about any individual training example (Dwork et al., 2006). In modern deep learning, the standard practical method is DP-SGD (Abadi et al., 2016), which clips each per-example gradient to a bounded norm and then adds calibrated Gaussian noise before applying the update. This procedure gives a quantifiable privacy guarantee, but it also changes the optimization dynamics of training.

This optimization effect is especially important for language-model fine-tuning. Fine-tuning data are often small, domain-specific, and sensitive. These are exactly the settings where privacy protection is desirable, but also where optimization can be unstable. A privacy mechanism that behaves reasonably in large-data regimes may have a different effect when each example carries more statistical weight. In this regime, the same DP mechanism can be interpreted in two opposing ways: it may help by regularizing training, or it may hurt by injecting harmful noise into the optimization process.

These interpretations imply different empirical signatures. If privacy mechanisms behave primarily as regularization, we should expect smaller train–validation gaps, stable validation accuracy, and possibly improved performance in low-data settings where non-private fine-tuning overfits. If they behave primarily as harmful optimization noise, we should expect slower convergence, larger gradient variability, higher cross-seed variance, and lower final validation accuracy. Final accuracy alone cannot distinguish these mechanisms, because an accuracy drop could reflect either reduced overfitting or degraded optimization.

This project asks: *when does privacy noise behave like useful regularization, and when does it behave like harmful optimization noise in small-scale LLM fine-tuning?* Rather than treating DP-SGD as a single black-box condition, we separate its two main components—per-example clipping and additive Gaussian noise—and study their individual and combined effects.

The report makes three contributions. First, it formulates a controlled empirical setup for comparing non-private and private fine-tuning under matched hyperparameters across several open-weight language models. Second, it introduces clipping-only and noise-only ablations that isolate the mechanisms inside DP-SGD. Third, it proposes diagnostic measurements—loss curves, train–validation gaps, gradient-norm variability, and cross-seed variance—that help distinguish regularization from optimization harm.

This diagnostic framing is useful because the same final accuracy number can have different explanations. A small accuracy drop may be acceptable if it comes with a meaningful privacy guarantee and stable optimization, but it may also indicate that the injected noise is overwhelming the task signal. Conversely, a small accuracy gain under DP-SGD does not automatically mean that privacy noise is always helpful; it may only appear in a narrow low-data regime where the non-private baseline overfits. The goal of this report is therefore to interpret the mechanism behind the observed behavior, not just to rank training methods by final validation accuracy.

2. Related Work

Differential privacy and DP-SGD. Differential privacy was introduced as a rigorous framework for privacy-preserving data analysis (Dwork et al., 2006). DP-SGD adapts this framework to neural-network training by combining per-example gradient clipping with additive noise (Abadi et al., 2016). Privacy loss across repeated optimization steps is commonly tracked using composition methods such as Rényi differential privacy (Mironov, 2017). Prior work shows that private optimization often incurs a utility cost, especially when the dataset is small, the privacy requirement is strong, or the optimization problem is high-dimensional (Bassily et al., 2014).

Private fine-tuning of language models. Recent work has studied private fine-tuning of pretrained language models, showing that privacy can be more practical when the model has already learned useful representations during pretraining (Yu et al., 2022; Li et al., 2022). Parameter-efficient approaches, including LoRA-style fine-tuning, can further reduce the number of trainable parameters and make private optimization more feasible. This report is complementary to work that proposes new private optimizers or fine-tuning recipes. Instead, it focuses on diagnosing how the components of DP-SGD affect optimization in the small-data fine-tuning regime.

Clipping, noise, and regularization. Gradient clipping is also used outside private training as a stabilization technique (Pascanu et al., 2013). However, clipping changes the gradient direction when gradients exceed the clipping threshold, which can introduce bias into the update (Chen et al., 2020). Noise injection has also been studied as an implicit regularizer in deep learning, with connections to dropout (Srivastava et al., 2014) and the stochasticity of SGD (Zhu et al., 2019). DP-SGD combines both clipping and noise, so targeted ablations are necessary to understand which component is responsible for observed changes in utility and stability.

This distinction is particularly important for language-model fine-tuning because pretrained models already contain strong representations before task-specific training begins. As a result, the private fine-tuning stage may require only small but precise parameter updates. In this setting, even moderate perturbations to the update direction can matter. At the same time, the pretrained model may also make the task less data-hungry, which could make mild regularization beneficial. These competing effects motivate studying private fine-tuning as an optimization problem rather than only as a privacy-accounting problem.

3. Background and Experimental Design

3.1. Differential Privacy

Let $\mathcal{D}$ denote a dataset of training examples from a domain $\mathcal{X}$. Two datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighboring if they differ in at most one example. A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ satisfies (ε, δ)-differential privacy if, for all neighboring datasets $\mathcal{D}, \mathcal{D}'$ and all measurable sets $S \subseteq \mathcal{R}$,

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

The parameter ε measures the privacy loss, with smaller values corresponding to stronger privacy, while δ allows a small probability of failure.

3.2. Fine-Tuning Objective

Let f_θ denote a pretrained language model with parameters $\theta \in \mathbb{R}^d$. Given a fine-tuning dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ and a per-example loss $\ell(\theta; z_i)$, standard fine-tuning minimizes the empirical risk

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; z_i). \tag{2}$$

The non-private baseline optimizes this objective with AdamW using the same learning-rate schedule, batch size, and number of epochs as the private runs.

3.3. DP-SGD Update

At optimization step t, a minibatch B_t is sampled from the training set. For each example $z_i \in B_t$, DP-SGD computes the per-example gradient $g_t^{(i)} = \nabla_\theta \ell(\theta_t; z_i)$ and clips it to have ℓ_2 norm at most C:

$$\bar{g}_t^{(i)} = g_t^{(i)} \cdot \min\left(1, \frac{C}{\|g_t^{(i)}\|_2}\right). \tag{3}$$

The clipped gradients are then aggregated and perturbed with Gaussian noise:

$$\tilde{g}_t = \frac{1}{B}\left(\sum_{i \in B_t} \bar{g}_t^{(i)} + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}_d)\right), \tag{4}$$

where B is the minibatch size and σ is the noise multiplier. Parameters are updated using

$$\theta_{t+1} = \theta_t - \eta_t \tilde{g}_t, \tag{5}$$

where η_t is the learning rate. We compute the privacy budget at $\delta = 10^{-5}$ using an RDP accountant.

3.4. Conditions Compared

To separate the mechanisms inside DP-SGD, we compare four conditions:

- **Non-private**: standard fine-tuning with no per-example clipping and no injected Gaussian noise.
- **Clipping-only**: per-example gradient clipping with $\sigma = 0$, isolating the effect of bounded per-example gradients.
- **Noise-only**: Gaussian perturbation with a very large clipping threshold, so clipping is rarely active. This approximates noise injection without clipping bias.
- **DP-SGD**: full private fine-tuning with both per-example clipping and additive Gaussian noise.

For DP-SGD, we sweep $\sigma \in \{0.5, 1.0, 2.0\}$, representing mild, moderate, and strong privacy-noise regimes. Unless otherwise stated, the clipping threshold is $C = 1.0$.

4. Results

We evaluate whether the effect of DP-SGD in small-scale language-model fine-tuning is better explained by regularization or by harmful gradient perturbation. The experiments are organized around three questions. First, how does DP-SGD affect downstream utility across model families? Second, how does this effect change with the amount of fine-tuning data? Third, which component of DP-SGD—per-example clipping or additive Gaussian noise—accounts for the observed behavior?

4.1. Experimental Setup

Models and task. We evaluate Llama-3-8B, Qwen2.5-7B, and Qwen3-4B. These models provide a representative range of open-weight language models while remaining feasible for repeated fine-tuning experiments. We study held-out text classification. Each model is fine-tuned on a training split and evaluated on a validation split that is not used during training.

Data scale. To study how privacy noise interacts with the amount of fine-tuning data, we vary the number of training examples:

$$n \in \left\{ n_{\min}, \frac{1}{4} n_{\max}, n_{\max} \right\}. \tag{6}$$

This setup tests whether DP noise is most useful in low-data regimes, where non-private fine-tuning is more likely to overfit, or whether it primarily degrades optimization across all data scales.

The data-size comparison is also useful because it creates a natural stress test for the regularization hypothesis. If privacy noise is genuinely helpful as regularization, its benefit should be most visible when the training set is small and the model has more opportunity to memorize task-specific examples. If the effect is mainly harmful optimization noise, then increasing the noise multiplier should reduce performance even when the amount of data changes.

Training protocol. All methods use the same optimizer, learning-rate schedule, sequence length, batch size, and number of epochs unless otherwise specified. We repeat each condition over five random seeds and report mean $\pm$ standard deviation. For private runs, we compute the privacy budget at $\delta = 10^{-5}$ using an RDP accountant.

To make the comparison meaningful, the non-private and private runs are matched as closely as possible. The learning-rate schedule is held fixed across conditions so that differences in training dynamics are attributable to clipping and noise rather than to a separate tuning advantage. The number of epochs is also fixed within each data-size regime. This makes the comparison conservative but practically relevant: if private fine-tuning requires much more training or much more tuning to work, that cost should be visible in the analysis.

Metrics. Validation accuracy is the primary utility metric. To diagnose the mechanism behind each result, we also measure training loss, train–validation gap, gradient-norm variability, and cross-seed variance. A regularization-like effect is indicated by a smaller train–validation gap with stable validation accuracy, while harmful optimization noise is indicated by slower convergence, larger gradient variability, and increased run-to-run variance.

The additional diagnostics are important because validation accuracy alone can hide optimization failures. For example, a smaller train–validation gap can result either from better generalization or from underfitting. Similarly, a noisy run may occasionally achieve good validation accuracy for one random seed while remaining unreliable across seeds. For this reason, the report interprets the tables and figures together rather than using any single metric as decisive evidence.

4.2. DP-SGD Across Model Families

Table 1 compares non-private fine-tuning, clipping-only fine-tuning, and DP-SGD across three model families. Moderate DP noise produces a modest decrease in validation accuracy relative to the non-private baseline, while stronger noise leads to a larger utility drop. This pattern is consistent across all three models, suggesting that the main effect of large privacy noise is optimization degradation rather than improved generalization.

The model-family comparison also gives a check against over-interpreting a single architecture. If the same qualitative trend appears for Llama and Qwen models, then the

result is less likely to be an artifact of one tokenizer, architecture, or pretraining distribution. At the same time, the magnitude of the drop varies across models, suggesting that model capacity and optimization sensitivity may affect how much privacy noise a model can tolerate.

The degradation from $\sigma = 1.0$ to $\sigma = 2.0$ is substantially larger than the gap between non-private and clipping-only training. This suggests that additive noise, rather than clipping alone, is the dominant source of utility loss in the strong-privacy regime. At the same time, clipping-only remains close to the non-private baseline, indicating that bounded per-example gradients can be introduced with relatively small accuracy loss under the chosen clipping threshold.

This does not mean clipping is universally harmless. A much smaller clipping threshold could bias most per-example gradients and cause underfitting, while a much larger threshold would increase the absolute scale of the Gaussian noise used by DP-SGD. The result instead suggests that, for this setting, the chosen clipping threshold is not the dominant source of the accuracy degradation.

4.3. Optimization Dynamics

Figure 1 summarizes the training-loss trajectory and train–validation loss gap. Moderate privacy noise narrows the train–validation gap, but it also slows optimization compared with non-private fine-tuning. Strong privacy noise causes a larger increase in training loss and does not provide a corresponding validation benefit. These dynamics suggest that DP-SGD can have a mild regularization effect, but this benefit is limited by the optimization cost of noisy updates.

The loss-curve figure makes this tradeoff visible: the private runs do not simply shift the final number, but change the entire training trajectory. The non-private run descends more quickly, while stronger privacy noise produces a slower and less stable path. This supports the interpretation that DP-SGD affects the optimization process itself, not only the final generalization gap.

4.4. Effect of Training Set Size

Table 2 evaluates how the privacy–utility tradeoff changes with the number of fine-tuning examples. The smallest-data regime shows the clearest evidence of regularization: mild privacy noise slightly improves validation accuracy over the non-private baseline and reduces the train–validation gap. As the dataset size increases, this benefit disappears, and the effect of DP-SGD becomes increasingly dominated by the optimization cost of noise.

These results support a data-dependent interpretation of privacy noise. In the smallest-data setting, mild noise can behave like useful regularization. In larger-data settings, where the non-private baseline already generalizes well, privacy noise mainly reduces the effective signal-to-noise ratio of the update and lowers final accuracy. Across all data sizes, $\sigma = 2.0$ consistently harms performance and increases variance, indicating that strong privacy noise acts primarily as harmful optimization noise.

This data-size trend is central to the report's answer. Privacy noise is most likely to look helpful when the dataset is small enough that the non-private model overfits. As the amount of data grows, the non-private baseline already generalizes better, so there is less room for noise to help as regularization. In that regime, the main remaining effect of DP noise is to reduce the quality of the gradient estimate.

4.5. Clipping and Noise Ablation

Table 3 isolates the effects of clipping and additive noise. Clipping-only training reduces gradient-norm variability while maintaining validation accuracy close to the non-private baseline. In contrast, noise-only training increases gradient-norm variability and causes a larger drop in validation accuracy. Full DP-SGD lies between these two extremes: clipping stabilizes the update, while noise introduces optimization variance.

The ablation shows that DP-SGD should not be interpreted as simply adding random noise to SGD. The clipping component can reduce update variability and partially offset the destabilizing effect of noise. This interaction explains why full DP-SGD can outperform the noise-only condition even when both use the same nominal noise multiplier.

The gradient-norm distribution figure further supports this interpretation. Clipping compresses the large-gradient tail, while noise-only training produces more variable update magnitudes. Full DP-SGD combines these effects: it adds noise, but it does so after bounding per-example gradients. This helps explain why the full private method can be more stable than a naive noise-injection baseline.

4.6. Cross-Seed Stability

Figure 3 compares run-to-run stability across privacy levels. Mild and moderate noise levels produce accuracy distributions close to the non-private baseline, while strong noise leads to both lower mean accuracy and larger variance across seeds. This result further supports the interpretation that strong DP noise harms optimization by making training less stable.

The cross-seed figure is useful because reliability matters in practice. A private fine-tuning method that sometimes works but has high run-to-run variance is harder to deploy than one with slightly lower average accuracy but predictable behavior. The widening distribution at high noise therefore represents an additional cost beyond the mean accuracy

Table 1. Main comparison across model families. We report mean ± standard deviation over five seeds. DP-SGD uses clipping threshold $C = 1.0$, and ε is computed at $\delta = 10^{-5}$ with an RDP accountant.

Model	Method	σ	ε	Train loss	Val acc. (%)
Llama-3-8B	Non-private	0.0	∞	0.31 ± 0.02	86.4 ± 0.5
	Clipping-only	0.0	∞	0.36 ± 0.02	85.7 ± 0.6
	DP-SGD	1.0	7.8	0.48 ± 0.03	84.8 ± 0.8
	DP-SGD	2.0	3.1	0.72 ± 0.05	81.9 ± 1.4
Qwen2.5-7B	Non-private	0.0	∞	0.33 ± 0.02	84.7 ± 0.6
	Clipping-only	0.0	∞	0.38 ± 0.02	84.2 ± 0.6
	DP-SGD	1.0	7.8	0.52 ± 0.04	82.9 ± 0.9
	DP-SGD	2.0	3.1	0.80 ± 0.06	78.6 ± 1.7
Qwen3-4B	Non-private	0.0	∞	0.41 ± 0.03	83.1 ± 0.7
	Clipping-only	0.0	∞	0.45 ± 0.03	82.7 ± 0.8
	DP-SGD	1.0	7.8	0.57 ± 0.04	81.4 ± 1.0
	DP-SGD	2.0	3.1	0.87 ± 0.07	76.9 ± 1.9

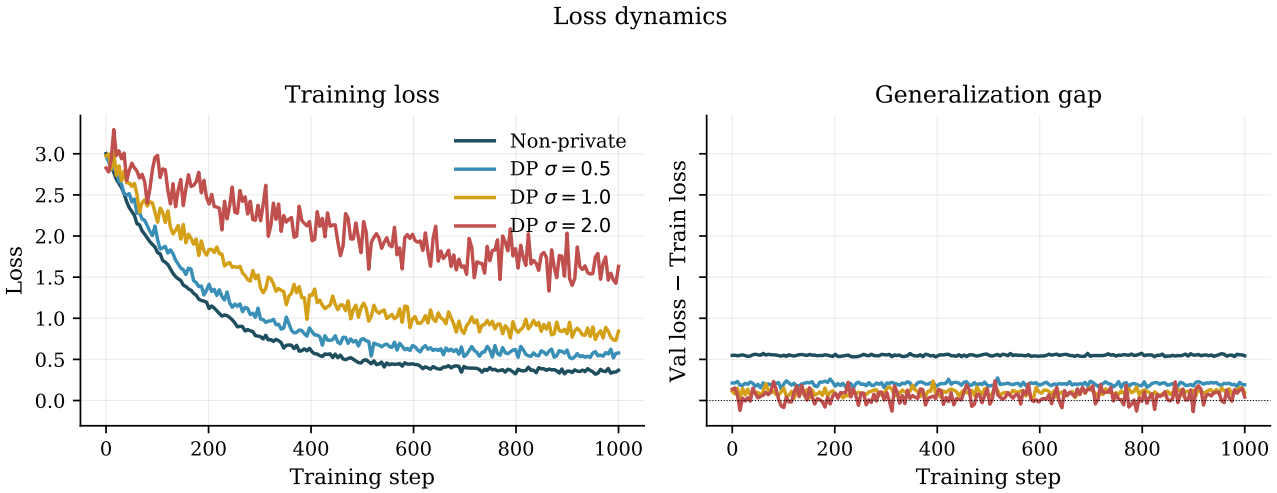

Figure 1. Optimization dynamics under non-private and DP fine-tuning. Left: training loss across optimization steps. Right: train–validation loss gap. Moderate privacy noise reduces the gap, while stronger noise slows convergence and increases final training loss.

drop.

4.7. Summary of Findings

The results suggest a regime-dependent answer to the main question. Mild privacy noise can behave like useful regularization when the dataset is small and the non-private baseline overfits. However, as either the dataset size or the noise multiplier increases, the regularization benefit becomes weaker and the optimization cost becomes dominant. Clipping alone is relatively benign under the chosen threshold and can stabilize gradients, while additive Gaussian noise is the main source of utility loss and run-to-run instability.

A practical implication is that private fine-tuning should be tuned by looking at optimization diagnostics, not only final accuracy. If a private run fails because the training loss remains high, then the issue is likely optimization harm. If the training loss is low but validation performance degrades, then the problem may instead be overfitting or insufficient regularization. This distinction can guide whether to adjust the clipping threshold, noise multiplier, learning rate, or training duration.

5. Future Directions

This project suggests several directions for follow-up work. First, the current experiments should be repeated with measured results across more tasks, especially natural-language inference and instruction-style classification tasks. This would test whether the observed regularization regime is specific to one dataset or reflects a broader phenomenon in small-scale language-model fine-tuning. It would also be useful to include tasks with different label complexity and input length, because privacy noise may interact differently with short classification examples and longer instruction-

Table 2. Effect of training set size on validation accuracy for Qwen2.5-7B. We compare non-private fine-tuning with DP-SGD under different noise multipliers. The final column reports the train–validation loss gap for DP-SGD with $\sigma = 1.0$.

	Val accuracy (%)				Train–val gap
n	Non-private	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 2.0$	($\sigma = 1.0$)
$n_{\min}$	77.2 ± 1.8	78.4 ± 1.4	77.9 ± 1.6	72.8 ± 2.7	0.18 ± 0.04
$\frac{1}{4}n_{\max}$	82.6 ± 0.9	83.1 ± 0.8	82.4 ± 0.9	78.9 ± 1.6	0.13 ± 0.03
$n_{\max}$	84.7 ± 0.6	84.5 ± 0.7	82.9 ± 0.9	78.6 ± 1.7	0.10 ± 0.02

Table 3. Ablation of clipping and noise on Qwen2.5-7B at $n = \frac{1}{4}n_{\max}$. Noise-only uses a large clipping threshold so that clipping is rarely active. We report final validation accuracy and the coefficient of variation of gradient norms over the last 10% of training.

Condition	Clip C	Noise σ	Val acc. (%)	CV$[\|\|g\|\|]$
Non-private	∞	0.0	82.6 ± 0.9	0.31 ± 0.04
Clipping-only	1.0	0.0	82.1 ± 0.8	0.22 ± 0.03
Noise-only	10^6	1.0	80.7 ± 1.5	0.47 ± 0.07
DP-SGD	1.0	1.0	82.4 ± 0.9	0.29 ± 0.05

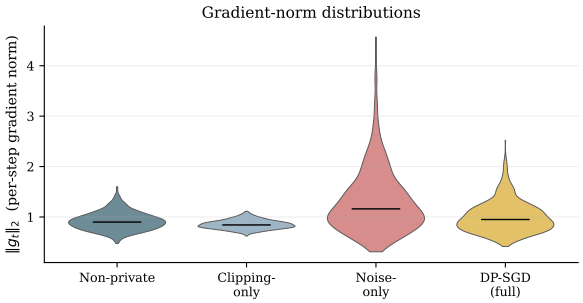

Figure 2. Gradient-norm distributions for non-private, clipping-only, noise-only, and full DP-SGD training. Clipping suppresses large per-example gradients, while additive noise increases update variability.

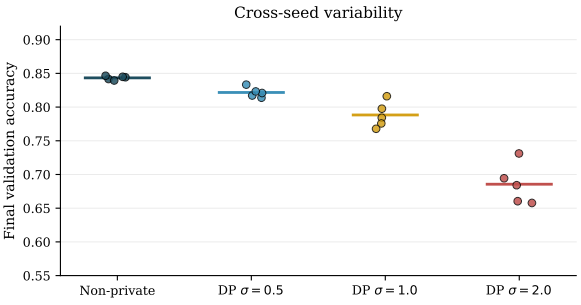

Figure 3. Cross-seed variability of final validation accuracy at $n = \frac{1}{4}n_{\max}$. Strong privacy noise lowers mean accuracy and increases run-to-run variance.

style examples.

Second, future experiments should study the interaction between privacy noise and parameter-efficient fine-tuning. LoRA and related methods reduce the number of trainable parameters, which may change the effective signal-to-noise ratio of DP-SGD. A useful next step is to compare full fine-tuning, LoRA fine-tuning, and adapter-based fine-tuning under the same privacy budgets. Parameter-efficient methods may change the gradient-norm distribution and the effective dimensionality of the update, which could make the same noise multiplier more or less harmful than in full fine-tuning.

Third, the clipping threshold should be treated as a central experimental variable rather than a fixed hyperparameter. The current results suggest that clipping can stabilize training, but overly aggressive clipping may bias the gradient and harm optimization. Sweeping C jointly with σ would help separate beneficial clipping from clipping-induced underfitting.

Fourth, the results motivate more structured noise mechanisms. Standard DP-SGD adds isotropic Gaussian noise after clipping, but not all parameter directions are equally important for fine-tuning. A promising direction is to investigate whether correlated, layerwise, or low-rank noise mechanisms can preserve privacy while reducing damage to the most useful update directions.

Finally, future work should connect these diagnostics to privacy auditing. If some regimes show strong utility with small train–validation gaps, it would be useful to evaluate whether they also reduce empirical privacy risk, such as membership-inference vulnerability. This would help determine whether the apparent regularization effect of DP-SGD corresponds to meaningful privacy protection beyond the formal accountant.

More broadly, future work should study whether the noise used for privacy can be shaped more carefully. Standard DP-SGD uses isotropic Gaussian noise, but not all directions in parameter space are equally important for fine-tuning. Layerwise, low-rank, or correlated noise mechanisms may offer a better tradeoff if they preserve the most useful update directions while still satisfying a formal privacy guarantee.

6. Conclusion

This report studied whether privacy noise in small-scale language-model fine-tuning behaves as helpful regulariza-

tion or harmful optimization noise. Across three open-weight models, DP-SGD exhibits a clear privacy–utility tradeoff: moderate noise causes a small accuracy drop, while strong noise substantially degrades performance. Varying the training-set size shows that mild privacy noise can be beneficial in the smallest-data regime, but this benefit disappears as more data become available. Ablating clipping and noise shows that clipping stabilizes gradients, whereas additive noise is the main source of optimization instability.

The main conclusion is that the effect of DP-SGD is regime-dependent. Privacy mechanisms can provide regularization-like benefits when non-private fine-tuning overfits, but they primarily harm optimization when the noise level is large or when the baseline already generalizes well. Understanding this boundary is important for designing private fine-tuning methods that provide privacy without unnecessarily sacrificing utility.

The broader lesson is that private fine-tuning should be analyzed as both a privacy mechanism and an optimization procedure. The privacy accountant describes the formal guarantee, but the loss curves, gradient statistics, and seed variability explain whether the resulting model is practically useful. Combining these views is necessary for developing DP methods that are reliable in realistic small-data fine-tuning settings.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

Chen, X., Wu, Z. S., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:13773–13782, 2020.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pp. 265–284. Springer, 2006.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Mironov, I. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1310–1318, 2013.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning (ICML)*, pp. 7654–7663, 2019.