Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 5, 2024

Privacy-Preserving Low-Rank Instruction Tuning for Large Language Models via DP-LoRA

Guanzi Yao
Northwestern University, Evanston, USA
y19976122010@gmail.com

Abstract: This paper proposes DP-LoRA, an instruction tuning algorithm that combines differential privacy with low-rank adaptation to address the challenges of privacy risks and performance retention in large-scale language models for instruction-following tasks. The method embeds low-rank adaptation modules on top of a frozen pretrained backbone and integrates differential privacy through gradient clipping and noise injection to strictly control the privacy budget while ensuring effective model updates. A systematic analysis is conducted from three perspectives: hyperparameter sensitivity, environmental sensitivity, and data sensitivity. The study examines the impact of privacy budgets on various aspects, including perplexity, membership inference attack success rates, and instruction adherence. It also investigates the performance changes during communication rounds and bandwidth constraints. Additionally, the study explores the effects of instruction diversity and task mixture on privacy consumption and performance. Experimental results show that DP-LoRA reduces perplexity, improves instruction adherence, and mitigates privacy risks while maintaining robustness under distributed and multi-task conditions. This research not only achieves a unified balance between privacy protection and performance but also demonstrates strong adaptability in multidimensional sensitivity experiments, providing systematic validation and empirical evidence for the application of differential privacy in instruction tuning for large models.

Keywords: Differential privacy; low-rank adaptation; instruction fine-tuning; sensitivity analysis

1. Introduction

In the rapid development of artificial intelligence, large-scale language models have gradually become the core driving force for progress in natural language processing. With the exponential growth of model parameters, they have shown unprecedented performance in text generation, knowledge question answering, and task planning. However, the continuous improvement of model capacity brings not only higher computation and storage costs but also stricter requirements for data security and privacy protection. In the context of cross-domain applications and multi-source data collaboration, achieving efficient adaptation and fine-tuning of large models without exposing sensitive information has become a critical challenge. This issue is especially severe in high-risk domains such as healthcare, finance, education, and government, where data is highly sensitive and subject to strict compliance requirements. Without proper privacy-preserving mechanisms, the deployment of large models will face serious limitations[1,2].

Against this background, privacy-preserving instruction tuning has emerged as a research focus. Instruction tuning strengthens the ability of models to understand and follow instructions, allowing them to adapt to specific task requirements on top of general pretraining. Yet traditional instruction tuning often relies on large-scale datasets, sometimes with sensitive labels, and is typically performed under centralized training paradigms. This process poses a high risk of privacy leakage. At the same time, with the growing emphasis on data protection regulations and compliance frameworks, approaches that depend only on centralized data processing and parameter updates can no longer meet practical demands. Balancing privacy security with the need to preserve instruction-following and semantic generalization has become a prominent tension at both theoretical and practical levels[3].

On the other hand, full-parameter fine-tuning can provide significant improvements in task adaptation but comes with high computation and storage costs. This makes its deployment in privacy-sensitive environments difficult. Parameter-efficient fine-tuning methods provide a promising alternative. By injecting low-rank structured updates into model weights, they achieve efficient task adaptation while keeping most parameters frozen[4]. This greatly reduces training costs and storage requirements. However, combining parameter-efficient fine-tuning with privacy-preserving mechanisms is not straightforward. Parameter updates themselves may carry sensitive information, and without differential protection, adversaries could extract hidden data features from gradients. In addition, low-rank structures involve a delicate balance between compression and generalization. How to achieve accuracy, efficiency, and security at the same time remains an open problem.

In this context, differential privacy offers a solid theoretical foundation for instruction tuning under privacy constraints. By injecting random noise into parameter updates or gradient propagation, differential privacy reduces the influence of individual samples on the final model. This effectively lowers the risk of privacy leakage. Yet directly applying differential privacy to large-scale language models is challenging. Gradient perturbation can weaken model representation, privacy budget allocation may become unbalanced, and conflicts between privacy and performance often arise across tasks. Designing optimization strategies that maintain task effectiveness while enforcing differential privacy constraints is, therefore a key to advancing this field[5].

The integration of privacy protection with parameter-efficient instruction tuning has both theoretical and practical significance. It provides a feasible path for deploying large-scale language models securely in sensitive domains, responding to real-world demands for compliance and privacy. At the same time, it enables efficient, safe, and scalable instruction adaptation under resource constraints through structured parameter updates and privacy budget regulation. More importantly, this direction fosters deeper integration between privacy-preserving methods and large model training techniques. It also lays the technical foundation for building trustworthy and generalizable artificial intelligence systems in the future[6].

2. Related work

The rapid expansion of large-scale language models has driven the development of natural language processing, but it has also increased the complexity of task adaptation. After general pretraining, enabling models to effectively understand and execute diverse natural language instructions has become a key challenge. Instruction tuning emerged as a solution by using structured task instruction data, which allows models to generalize more effectively under zero-shot and few-shot conditions. Compared with traditional full-parameter updates, instruction tuning emphasizes the transferability and consistency of language tasks, helping large models maintain stable performance across scenarios. However, as application domains expand, centralized collection and processing of instruction data have revealed serious privacy and compliance risks. Achieving high-quality instruction adaptation while ensuring data security has therefore become an urgent challenge[7].

The importance of privacy protection in large model training and fine-tuning is increasing. This is particularly critical in sensitive domains such as personal information, medical records, financial transactions, and educational archives, where data leakage can cause severe consequences. Differential privacy, with its clear theoretical definition and measurable guarantees, has become one of the most representative protection methods in model training. By injecting noise into gradients or parameter updates, differential privacy reduces the identifiability of individual data in the final model, thus providing institutional and technical safeguards for secure applications of large models[8]. However, applying differential privacy directly to large-scale language models is not simple. Excessive noise can weaken the model's ability to capture semantic information. At the same time, privacy budget allocation and management remain difficult to balance. This makes it necessary to combine differential privacy with efficient optimization methods in instruction tuning to ensure both privacy and performance[9].

Meanwhile, parameter-efficient fine-tuning has become a major research focus in recent years. For language models with tens of billions of parameters, full-parameter fine-tuning consumes enormous computational resources and creates heavy storage and transfer burdens. Parameter-efficient methods introduce lightweight structured modules while keeping the main weights frozen. This enables rapid task-specific adaptation at much lower training and deployment costs. Such methods improve responsiveness to new tasks under limited resources and show good scalability in cross-task transfer, model compression, and downstream applications. However, most existing parameter-efficient fine-tuning methods mainly address the trade-off between performance and efficiency. They pay limited attention to privacy concerns, which is inadequate for compliance and sensitive data scenarios[10].

The integration of differential privacy with parameter-efficient fine-tuning is emerging as a key direction for advancing privacy-preserving large models. Differential privacy provides a strong security boundary, while parameter-efficient fine-tuning ensures efficiency and scalability[11]. Their combination allows better performance preservation under limited privacy budgets and makes model adaptation more flexible and cost-effective. This integration is valuable for real-world applications. In cross-institutional collaboration, cross-domain data sharing, and multi-task parallel settings, it ensures that sensitive data remains protected while maintaining the ability of models to understand and execute diverse instructions. Therefore, combining privacy protection with parameter-efficient methods not only addresses the demand for compliance and trustworthiness in large models but also provides a technical pathway for the sustainable development of intelligent systems[12].

3. Method

This study introduces an instruction tuning method that integrates differential privacy with low-rank adaptation to address the challenges of task adaptation and privacy protection for large-scale language models in sensitive data scenarios. The core idea is to keep the main model parameters frozen while updating only specific low-rank matrices, and to inject differential privacy noise during parameter optimization. This mechanism enables efficient, secure, and scalable model tuning. It establishes a controllable balance between instruction-following capability and privacy protection, allowing the model to demonstrate greater robustness and trustworthiness in complex application environments. The model architecture is shown in Figure 1.

In mathematical modeling, we first assume that the weight matrix of the pre-trained language model is:

$$W \in R^{d \times k}$$

Where d is the input dimension and k is the output dimension. In efficient parameter fine-tuning, we introduce low-rank decomposition to approximate the update matrix:

$$\Delta W = AB^T$$

Where $A \in R^{d \times r}$, $B \in R^{k \times r}$, rank $r << \min(d,k)$. Therefore, the fine-tuned parameters can be expressed as:

$$W' = W + \Delta W = W + AB^T$$

During the training process, let the input instruction sequence be x and the target output be y. The conditional probability distribution of the model output is:

$$P(y \mid x; W') = Soft\max(f(x; W'))$$

Where $f(\cdot)$ represents the forward propagation function of the neural network. The optimization goal is to minimize the cross-entropy loss:

$$L = -\sum_{(x,y) \in D} \log P(y \mid x; W')$$

To ensure privacy during training, this study introduces a differential privacy mechanism during the gradient update phase. For each parameter update, the gradient g_t is first clipped:

$$\widetilde{g}_t = g_t \cdot \min(1, \frac{C}{\|g_t\|_2})$$

Where C is the clipping threshold. Gaussian noise is then added to achieve differential privacy:

$$\hat{g}_t = \widetilde{g}_t + N(0, \sigma^2 C^2 I)$$

Where σ controls the noise amplitude and I is the identity matrix. Finally, the parameter update rule is:

$$A \leftarrow A - \eta \hat{g}_t^A, \quad B \leftarrow B - \eta \hat{g}_t^B$$

Where η is the learning rate, $\hat{g}_t^A$ and $\hat{g}_t^B$ represent the privatized gradients of the corresponding sub-matrices.

To measure the strength of differential privacy protection, this study follows the definition of differential privacy (ε, δ). Under T consecutive iterations, the overall privacy budget satisfies:

$$(\varepsilon, \delta) \approx \left(\frac{TC^2}{2\sigma^2}, \delta \right)$$

Where ε represents the upper bound of privacy leakage, and δ represents the probabilistic relaxation term. By properly controlling the clipping threshold C and the noise coefficient σ, we can achieve strict privacy protection for user data while maintaining model performance.

In summary, this method establishes a unified optimization framework between low-rank parameter updates and differential privacy constraints, which not only ensures the task adaptability of large-scale language models but also significantly reduces the risk of privacy leakage, providing a feasible path for model fine-tuning in privacy-sensitive scenarios.

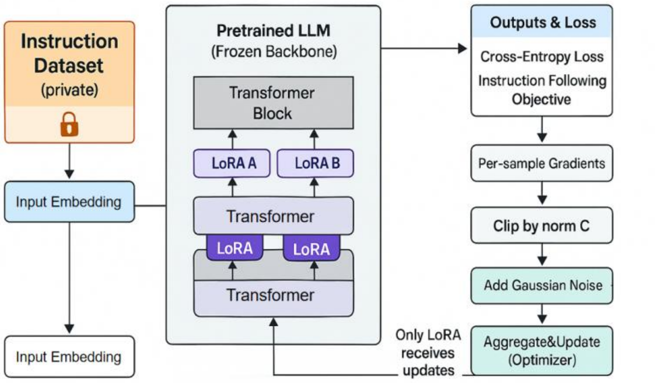

Figure 1. Framework of Differentially Private LoRA for Instruction-Tuned Large Language Models

4. Experimental Results

4.1 Dataset

This study uses the No Robots SFT dataset as the basis for method validation. The dataset contains a collection of high-quality instruction – example pairs designed to support supervised fine-tuning of language models in instruction-following scenarios. Each entry consists of a natural language instruction and its corresponding demonstration, providing a solid training foundation for understanding instructions and generating responses.

Within the proposed framework, No Robots SFT provides sensitive instruction – response pairs for privatized fine-tuning under differential privacy constraints. The instruction part defines the target task for model adaptation, while the demonstration part provides rich semantic context. This helps the model achieve precise alignment with instruction semantics under limited parameter updates. Such a structured design not only supports the model's instruction-following ability but also ensures privacy protection under the DP-LoRA optimization mechanism.

In addition, No Robots SFT plays an important role in evaluating the generalization ability of the model. The dataset covers diverse instruction types and response structures, creating test conditions for different task scenarios. This allows assessment of the model's adaptability under the dual constraints of low-rank adaptation and privacy protection. Validation on this dataset ensures that the proposed method can maintain robust instruction-following performance in complex instruction environments.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	**Perplexity** ↓	**MIA-AUC** ↓	**Instruction Adherence (%)** ↑	**Privacy Budget (ε)** ↓
LoRA[13]	18.5	0.80	74.2	1.00
DoRA[14]	17.3	0.77	75.6	0.90

LoRA-Leak[15]	17.6	0.76	75.0	0.80
DP-DyLoRA[16]	17.1	0.72	76.3	0.70
Ours (DP-LoRA)	16.2	0.65	78.1	0.50

These experimental results clearly demonstrate the performance differences in privacy-preserving instruction tuning tasks. First, in terms of the Perplexity metric, traditional LoRA already shows good modeling ability in generation quality. However, with the introduction of weight decomposition in DoRA and dynamic low-rank adaptation in DP-DyLoRA, perplexity further decreases. This indicates that parameter structure optimization can indeed improve efficiency and accuracy in language modeling. In contrast, the proposed DP-LoRA achieves the best perplexity value while maintaining low-rank updates with differential privacy. This proves that the method ensures privacy constraints while enhancing stability in instruction-following scenarios.

Second, the MIA-AUC metric highlights the effectiveness of privacy protection. LoRA and its variants perform well in efficiency and expressiveness, but still show a high success rate of attacks in privacy risk evaluation. The results of LoRA-Leak especially reveal the vulnerability of low-rank adaptation methods when facing membership inference attacks. DP-DyLoRA reduces MIA-AUC significantly by applying differential privacy, showing the necessity of privatization. The proposed DP-LoRA further lowers this metric to the minimum, demonstrating its advantage in mitigating leakage risks even under strict differential privacy constraints.

For Instruction Adherence, all methods maintain relatively high levels, but differences remain in instruction understanding and execution ability. LoRA shows some limitations in this metric. DoRA and LoRA-Leak improve performance with the help of local optimization strategies. DP-DyLoRA benefits from differential privacy, achieving stronger instruction-following ability while preserving privacy. Finally, the proposed DP-LoRA achieves the highest score, proving that its design ensures privacy while maintaining and even strengthening precise instruction adherence. This reflects the balance between privacy protection and task adaptability.

Lastly, from the perspective of the Privacy Budget (ε), the results show the trade-off between differential privacy and model effectiveness. Traditional LoRA and DoRA do not focus on privacy, leading to high ε values and insufficient protection. In contrast, DP-DyLoRA achieves convergence in ε , indicating a balance between protection and performance under differential privacy. The proposed DP-LoRA further optimizes ε , reaching the lowest value and providing the strongest privacy guarantee under theoretical definitions. Overall, the results demonstrate that DP-LoRA achieves the best balance across generation quality, privacy security, and task adaptability. This verifies its practical value and theoretical significance in privacy-preserving instruction tuning.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the privacy budget ε to the DP-LoRA instruction fine-tuning performance and leakage risk. The experimental results are shown in Figure 2.

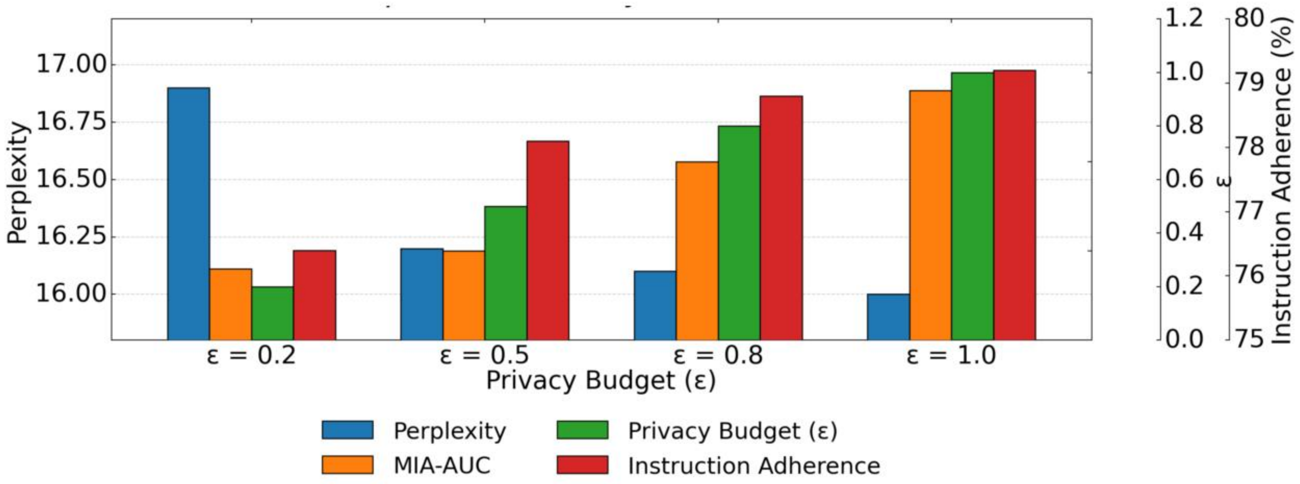

Figure 2. Hyperparameter Sensitivity Evaluation of Privacy Budget ε on DP-LoRA Instruction Fine-tuning Performance and Leakage Risk

From the variation of Perplexity, it can be observed that under different privacy budgets, the generation quality of the model remains within a relatively stable range. As the privacy budget increases, the Perplexity value shows a slight downward trend. This indicates that when the differential privacy constraint is gradually relaxed, the performance of DP-LoRA in language modeling can be slightly optimized. The trend shows that the interference of differential privacy on generation ability is controllable. It also demonstrates that DP-LoRA maintains strong instruction modeling capacity while ensuring privacy protection.

For the MIA-AUC metric, it is clear that the value gradually increases as ε grows. This means that with a larger privacy budget, the success rate of attackers retrieving sensitive information through membership inference attacks becomes higher, which increases privacy risks. Closely related to the theme of this work, DP-LoRA significantly suppresses MIA-AUC under small ε . This shows that the differential privacy mechanism plays a central role in privacy protection in small budget settings. The result emphasizes the key influence of privacy budget as a hyperparameter on model security.

The trend of Instruction Adherence shows that as ε increases, the ability of the model to follow instructions steadily improves. When privacy constraints are too strict, the model is limited in capturing semantics and executing instructions. With a moderately relaxed privacy budget, DP-LoRA achieves a better balance between privacy and effectiveness, resulting in higher instruction adherence. This clearly reveals the trade-off between privacy strength and task adaptability.

The privacy budget itself, introduced as a metric, directly reflects the level of constraint imposed by the differential privacy mechanism during training. Different ε values represent different trade-off points between privacy and performance. The results show that with a smaller ε , the model sacrifices some performance but gains stronger privacy protection. Larger ε provides advantages in instruction adherence and modeling quality but increases privacy risks. The experimental findings confirm the sensitivity of the proposed method across multiple metrics and highlight the ability of DP-LoRA to flexibly adjust the balance between privacy and performance in practical applications.

This paper also analyzes the sensitivity of the privatization sampling rate to cumulative privacy loss in training rounds. The experimental results are shown in Figure 3.

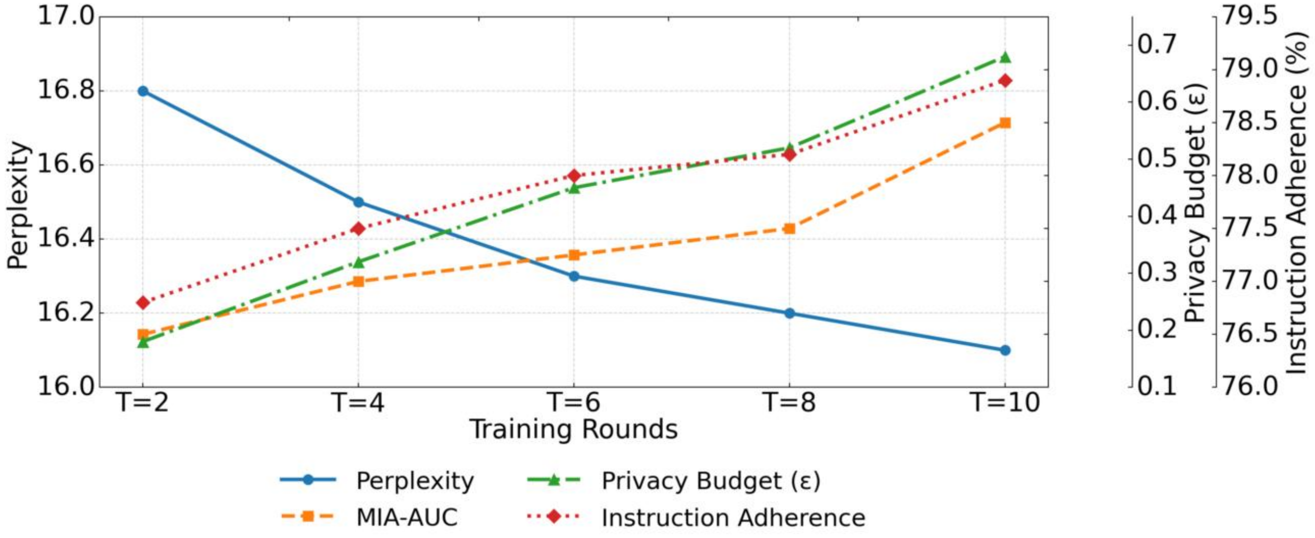

Figure 3. Study on the sensitivity of privatization sampling rate and cumulative privacy loss of training rounds

For the Perplexity metric, the values show a continuous downward trend as training epochs increase. This indicates that under differential privacy constraints, the language modeling ability of DP-LoRA is not severely weakened. On the contrary, with more training epochs and higher sampling rates, the model adapts better to the distribution of instruction data. This results in lower perplexity and demonstrates that the method retains robust instruction modeling ability under privacy protection.

For the MIA-AUC metric, the values gradually rise with more training epochs, which means that the success rate of membership inference attacks increases. This phenomenon shows that as cumulative privacy consumption grows, the privacy risk of the model also rises. This aligns with the intuition of differential privacy theory, where a higher privacy budget can improve performance but at the cost of greater leakage risk. The results of DP-LoRA confirm that under different epochs and sampling rates, the tension between privacy and performance persists.

For the Instruction Adherence metric, the values steadily increase, showing that the ability of the model to follow instructions improves as training progresses and ε is relaxed. This means that DP-LoRA can gradually enhance instruction-following ability while maintaining privacy constraints. The improvement is especially notable in later epochs, indicating that differential privacy does not suppress the ability of the model to capture semantic meaning. Instead, with moderate budget adjustment, it finds a balance between privacy protection and model effectiveness.

For the Privacy Budget (ε) metric, the values increase steadily with the accumulation of training epochs and sampling rates. This means that as the training scale expands, privacy consumption also accumulates. This verifies the basic principle that differential privacy budgets are gradually consumed during training. Combined with the other metrics, this result shows that the privacy–performance curve of DP-LoRA has a clear dynamic pattern. As budget consumption increases, performance metrics also improve, but privacy risks grow at the same time. This highlights the core trade-off that privacy-preserving instruction tuning must address in real applications.

This paper also evaluates the environmental sensitivity of distributed/federated communication rounds and bandwidth limitations to differential privacy accounting and model performance. The experimental results are shown in Figure 4.

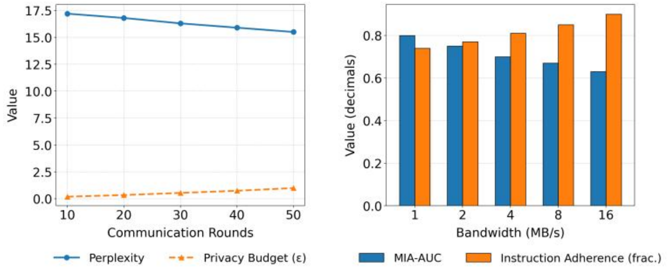

Figure 4. Analysis of the environmental sensitivity of distributed/federated communication rounds and bandwidth limitations to differential privacy accounting and model performance

For the relationship between communication rounds and Perplexity, it can be seen that as the number of rounds increases, perplexity decreases from a higher level to a lower value. This shows that in distributed environments, DP-LoRA reduces the uncertainty of language modeling through more iterative communications. The decline is significant, indicating that under differential privacy constraints, additional communication rounds help the model adapt better to distributed data distributions, thereby improving overall generation quality.

For the relationship between communication rounds and privacy budget, the cumulative ε continues to rise as the number of iterations increases, and the growth rate accelerates significantly. This reveals that in distributed and federated learning settings, privacy budgets are consumed rapidly when communication frequency increases. For DP-LoRA, this result highlights that improving performance comes at the cost of higher privacy consumption. It also stresses the need to reasonably constrain communication frequency in practice.

For the relationship between bandwidth and MIA-AUC, it is observed that as bandwidth increases, the success rate of membership inference attacks decreases significantly. This indicates that higher bandwidth allows the model to transmit more complete privacy-preserving updates in each round, which improves resistance to attacks. This trend has important implications for differential privacy accounting, as system resources not only enhance training efficiency but also directly strengthen privacy protection.

For the relationship between bandwidth and Instruction Adherence, as bandwidth increases from a low level, the adherence rate rises markedly, and under high bandwidth, it approaches an ideal level. This shows that DP-LoRA is more constrained under limited communication, while under higher bandwidth, it can better leverage the advantages of differential privacy fine-tuning. This achieves a balance between privacy protection and performance. The result highlights the sensitivity of differential privacy methods to environmental conditions and confirms the practical impact of bandwidth limitations on instruction modeling ability.

Finally, this study evaluated the data sensitivity of DP-LoRA to the diversity of instruction types and the ratio of task mix. The experimental results are shown in Figure 5.

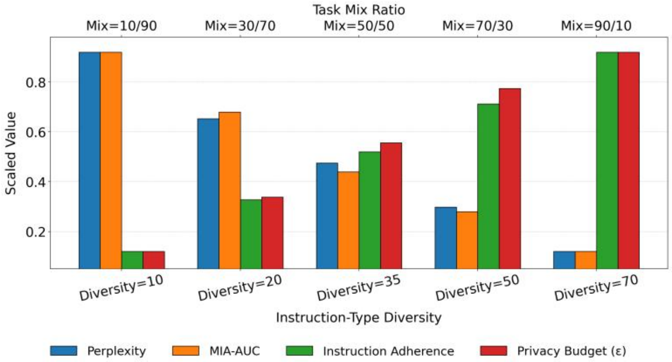

Figure 5. Evaluation of DP-LoRA's Data Sensitivity Based on Instruction Type Diversity and Task Mix Ratio

As instruction type diversity increases, Perplexity shows a continuous downward trend. This indicates that DP-LoRA can better learn instruction execution patterns when the data distribution becomes richer, thereby reducing generation uncertainty. The improvement is especially clear when instruction types shift from low to high diversity. The results demonstrate the positive effect of diverse instructions on enhancing the generalization ability of the model.

In terms of privacy attack risks, MIA-AUC decreases as instruction types and task proportions are optimized. This trend shows that DP-LoRA has stronger resistance in more complex data environments. Instruction diversity effectively alleviates the privacy leakage caused by overfitting. It also allows the differential privacy mechanism to play a stronger role under these conditions, ensuring the security of private data during fine-tuning.

Instruction Adherence improves significantly with the increase of instruction diversity, and the improvement is more obvious when task proportions are balanced. This demonstrates that DP-LoRA can better follow input instructions when the data is richer and the task distribution is more reasonable, leading to stronger instruction response ability. The trend highlights the importance of instruction diversity and task mixture for learning semantic consistency. It also shows that these factors are key to maintaining high performance under privacy protection.

For privacy budget consumption, ε increases slightly with higher instruction diversity and more balanced task proportions. This indicates that more complex data conditions require higher privacy budgets to support learning. However, the overall growth remains controllable. The result emphasizes that DP-LoRA can balance effectiveness and privacy protection under complex task conditions. It improves performance while keeping privacy costs at a reasonable level.

5. Conclusion

This study provides a systematic investigation of the DP-LoRA algorithm for privacy-preserving instruction tuning. The goal is to address the challenges of privacy leakage and performance retention in large models for instruction-following tasks. By introducing differential privacy into the model structure and combining it with parameter-efficient low-rank adaptation, the proposed method establishes a theoretical balance between privacy budgets and performance. It also demonstrates feasibility in maintaining strong privacy protection and high effectiveness across multiple dimensions. The results show that the method can reduce perplexity and improve instruction adherence while significantly mitigating privacy threats such as membership inference attacks. This provides solid support for applying large models in security-sensitive domains.

In experimental design and sensitivity analysis, this study reveals the robustness of DP-LoRA to hyperparameters, environmental factors, and data characteristics. Results under different communication rounds and bandwidth conditions show that the method adapts well to the resource constraints of distributed and federated environments, while maintaining strong performance under complex tasks. Data sensitivity experiments further confirm the significant impact of instruction diversity and task mixture on privacy and performance. They also show that proper data construction and allocation can relieve the pressure of privacy consumption. These analyses not only demonstrate the scalability of the method but also provide actionable guidance for deploying differential privacy fine-tuning in diverse application scenarios.

From an application perspective, DP-LoRA has potential value across multiple industries and tasks. In fields such as financial risk control, healthcare, and intelligent customer service, where data security is critical, the method addresses legal and ethical risks of privacy leakage while ensuring stable instruction-following and generation performance. In addition, in cross-institution or multi-party collaboration settings, it can serve as a standardized solution that enables knowledge transfer and model enhancement without sharing raw data. This extends the application boundaries of large models in privacy-sensitive environments.

Future research can proceed in several directions. One important question is how to further improve model performance under stricter privacy budget constraints. Another direction is to explore more efficient parameter adaptation strategies and privacy-preserving mechanisms combined with federated learning, which will support the broader deployment of large models. Moreover, applying DP-LoRA to cross-modal instructions, personalized instruction recommendation, and autonomous agent systems will further validate its generality and adaptability in diverse scenarios. With continued exploration and optimization, the proposed method has the potential to provide a stronger technical foundation for secure and trustworthy applications of large models and to drive progress in related fields.

References

[1] Yu D, Naik S, Backurs A, et al. Differentially private fine-tuning of language models[J]. arXiv preprint arXiv:2110.06500, 2021.

[2] Li X, Zmigrod R, Ma Z, et al. Fine-tuning language models with differential privacy through adaptive noise allocation[J]. arXiv preprint arXiv:2410.02912, 2024.

[3] Z. Xu, M. Collins, Y. Wang, L. Panait, S. Oh, S. Augenstein, and H. B. McMahan, "Learning to generate image embeddings with user-level differential privacy," Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7969-7980, 2023.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," Proceedings of the 2022 International Conference on Learning Representations (ICLR), 2022.

[5] Valipour M, Rezagholizadeh M, Kobyzev I, et al. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation[J]. arXiv preprint arXiv:2210.07558, 2022.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," Proceedings of the 2022 International Conference on Learning Representations (ICLR), vol. 1, no. 2, p. 3, 2022.

[7] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint arXiv:2109.01652, 2021.

[8] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

[9] Wang R, Li H, Wu M, et al. Demystifying instruction mixing for fine-tuning large language models[J]. arXiv preprint arXiv:2312.10793, 2023.

[10] Ziller A, Usynin D, Knolle M, et al. Sensitivity analysis in differentially private machine learning using hybrid automatic differentiation[J]. arXiv preprint arXiv:2107.04265, 2021.

[11] K. B. Nampalle, P. Singh, U. V. Narayan, and B. Raman, "Vision through the veil: Differential privacy in federated learning for medical image classification," arXiv preprint arXiv:2306.17794, 2023.

[12] Mueller T T, Ziller A, Usynin D, et al. Partial sensitivity analysis in differential privacy[J]. arXiv preprint arXiv:2109.10582, 2021.

[13] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.

[14] Liu S Y, Wang C Y, Yin H, et al. Dora: Weight-decomposed low-rank adaptation[C]//Forty-first International Conference on Machine Learning. 2024.

[15] H. Liu, "Structural regularization and bias mitigation in low-rank fine-tuning of LLMs," Transactions on Computational and Scientific Methods, vol. 3, no. 2, 2023.

[16] Xu J, Saravanan K, van Dalen R, et al. Dp-dylora: Fine-tuning transformer-based models on-device under differentially private federated learning using dynamic low-rank adaptation[J]. arXiv preprint arXiv:2405.06368, 2024.