SCIENCE CHINA
Information Sciences

• LETTER •

July 2026, Vol. 69, Iss. 7, 179104:1–179104:2
https://doi.org/10.1007/s11432-025-4859-8

Resource-efficient federated fine-tuning for privacy-preserving NLP applications

Dongqi CAI[1,3], Shangguang WANG[1], Yaozong WU[1], Hanlin GU[2], Yifan DUAN[1], Lixin FAN[2], Nicholas D. LANE[3] & Mengwei XU[1*]

[1]*Department of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*WeBank, Shenzhen 518000, China*
[3]*Department of Computer Science, University of Cambridge, Cambridge CB2 1TN, UK*

Received 18 February 2025/Revised 28 July 2025/Accepted 11 March 2026/Published online 8 June 2026

Citation Cai D Q, Wang S G, Wu Y Z, et al. Resource-efficient federated fine-tuning for privacy-preserving NLP applications. Sci China Inf Sci, 2026, 69(7): 179104, https://doi.org/10.1007/s11432-025-4859-8

Transformer architectures and their variants have propelled natural language processing (NLP) models into practical use for on-device scenarios. An example is Android AICore, a system-level service already integrated into several Google applications to deliver features like summarization and auto-reply directly on user devices.

A major driver behind the success of modern NLP models is their ability to learn from massive datasets. The standard pipeline consists of two phases: pre-training, which extracts general linguistic patterns from large public corpora, and fine-tuning, which tailors the model to specific downstream services. The latter stage frequently involves user-generated, private data that reside across mobile devices.

To enable fine-tuning on such decentralized and sensitive datasets, federated learning (FL) has become the prevailing solution [1]. In this paradigm, a central server coordinates mobile devices, each of which trains a local model replica using its private data. Only the model updates—not raw data—are exchanged: clients periodically send their updates to the cloud, which aggregates them and redistributes the improved model. This cycle is repeated until the model converges, often requiring hundreds or even thousands of rounds. To mitigate privacy risks such as gradient leakage, differential privacy (DP) techniques are commonly applied [2].

This work concentrates on the privacy-preserving fine-tuning of NLP models in federated environments, a process widely referred to as FedNLP. Empirical study, however, reveals that FedNLP suffers from severe training delays due to large model sizes and the resulting network and computational burden. To address this, our prior work FedAdapter [1] focuses on adapters—compact bottleneck modules inserted at selected layers—as a key enabler of efficient training. The challenge lies in determining the right depth and width of these adapters, which critically affects performance and varies across tasks, accuracy goals, and device capacities. Instead of relying on fixed settings, FedAdapter proposes a dynamic framework designed to make FedNLP practical through two core innovations. First, FedAdapter adopts progressive configuration: it begins with lightweight adapters at shallow layers to capture basic patterns, and gradually expands deeper and wider adapters to learn more complex representations. Second, FedAdapter continuously profiles upcoming configurations by assigning participants to lightweight trials, enabling informed adjustments during training. Extensive evaluations confirm that FedAdapter reduces FedNLP's convergence time to within a few hours—achieving up to 155.5× speedup over the baseline and 48× over strong alternatives—demonstrating the effectiveness of our method.

Atop FedAdapter [1], this work further addresses privacy concerns with an adaptive differential privacy scheme. In some scenarios, even federated learning may fall short in safeguarding privacy, as models can remain vulnerable to privacy leakage. For instance, active adversaries can infer membership information from the uploaded adapter weights, revealing whether specific data points were included in the training set. Recent studies indicate that fine-tuned large language models are prone to memorizing user data. For example, Ref. [3] demonstrated that membership inference attacks on adapter-tuned models can recover over 15% of the training data. Fine-tuning only the head layer, i.e., the last layer of the model, is even more susceptible, showing data leakage rates as high as 81.6%. Because the head layer is where the next word prediction happens. Numerical analyses reveal that implementing strong user-level DP guarantees significantly reduces unintended memorization [4]. This approach provides a robust solution for minimizing privacy risks in federated fine-tuning scenarios.

However, DP may negatively impact model performance. This decline is mainly attributed to catastrophic noise accumulation. The effect of the injected noise becomes increasingly severe as training progresses. For example, by the 10th round of training, the model's accuracy drops drastically to 2%, compared to 72% achieved in the second round.

To address the accuracy degradation introduced by DP while preserving privacy protection, our method employs an adaptive noise injection strategy from both layer-wise and iteration-wise perspectives. (i) Noise is injected only into the updated adapter layers in the upper layers, which are the sole trainable and trans-

* Corresponding author (email: mwx@bupt.edu.cn)

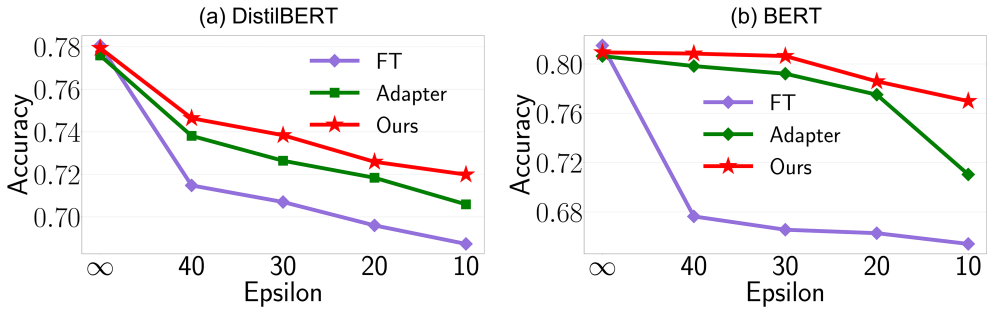

Figure 1 (Color online) Impact of DP on accuracy performance with different training methods.

ferable parameters during federated fine-tuning. Limiting updates to a small subset of model weights reduces noise accumulation, a finding supported by prior work [2]. The lower layers, being closer to raw data and hence more susceptible to privacy leakage, remain largely frozen during fine-tuning. (ii) To mitigate temporal gradient explosion risks in early training stages [5], which may intensify privacy leakage, training initially updates only a limited number of top-layer adapters. As training proceeds, more adapters are introduced, and the number of local iterations is adjusted dynamically to optimize privacy budget usage. Specifically, during the training, the number of local iterations matches the number of globally tunable adapters. This adaptive DP strategy is integrated with our auto-configurator that adjusts adapter configurations based on the training phase, balancing accuracy and privacy.

To formalize this adaptive privacy control, we allocate the total user-level privacy budget $\varepsilon_{\text{total}}$ across training rounds proportionally to the number of updated adapters and local iterations per round, which jointly serve as a proxy for per-round privacy cost. In round t, the allocated budget ε_t is calculated as

$$\varepsilon_t = \varepsilon_{\text{total}} \cdot \frac{a_t \cdot L_t}{\sum_{j=1}^{T} a_j \cdot L_j},$$

where a_t is the number of adapter layers updated in round t, and L_t denotes the number of local iterations. This allocation emphasizes later rounds with more informative updates, while early rounds consume less budget. Such dynamic adjustment maintains privacy guarantees without prematurely exhausting the budget, sustaining model utility throughout training.

During the privacy-preserving federated fine-tuning process, only the adapter parameters are transmitted. Noise is injected into these parameters using Gaussian mechanisms, consistent with established methods for ensuring differential privacy [2], thereby maintaining the same level of privacy guarantees as prior approaches. Attributed to the parameter-efficient adapter fine-tuning, where only a small portion of the model is updated, the additional computational cost of calculating global DP sensitivity remains minimal compared to the cost of gradient updates.

We evaluate the influence of DP across various methods and privacy budgets ε, as shown in Figure 1. We use BERT as representative models. BERT has 12 transformer layers; DistilBERT has 6 and achieves similar performance with 40% fewer parameters and 60% faster inference. Pre-trained weights are sourced from Hugging Face. Our method is evaluated on AGNEWS, a non-IID dataset for text classification. Default settings follow FedAdapter [1]: batch size = 4, learning rate = 0.1, sequence length = 256 (or 64 where specified). Each round selects 15 clients. For DP evaluation: δ = 1e−5, sensitivity = 0.1, clip norm = 10, client query budget = 50. We compare our method to (1) `FT`: full-model fine-tuning on each client and (2) `Adapter`: all adapter layers are updated and subjected to noise injection. All use the same aggregation (FedAvg) and client sampling for fair comparison.

Adapter demonstrates greater resilience to higher levels of noise injection due to its parameter-efficient design, where only a small subset of model parameters are updated [2]. This advantage is particularly evident in larger models, where `Adapter` improves convergence performance by up to 12.4% compared to `FT`. Our method achieves a superior trade-off between privacy and utility across all evaluated models and datasets. This improvement is attributed to the adaptive noise injection strategy, which is orchestrated through sideline adapter configuration trials.

Our method presents a novel federated learning framework designed for efficient fine-tuning of NLP models. By leveraging adapter modules as the sole trainable components, our method substantially lowers computation and communication overhead, while also enhancing the balance between privacy and model utility. To dynamically discover effective adapter configurations during training, our method employs a progressive optimization strategy coupled with a trial-and-error exploration mechanism. To preserve model utility under privacy constraints, our method also integrates an adaptive differential privacy mechanism. Extensive evaluations demonstrate that our method consistently achieves significant acceleration in training compared to prior state-of-the-art methods.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62425203), Royal Academy of Engineering via DANTE (a RAEng Chair), European Research Council, specifically the REDIAL Project, SPRIND under the Composite Learning Challenge, and Google through a Google Academic Research Award.

References

1 Cai D, Wu Y, Wang S, et al. Efficient federated learning for modern NLP. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023. 1–16
2 Zhao H, Du W, Li F, et al. FedPrompt: communication-efficient and privacy-preserving prompt tuning in federated learning. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. 1–5
3 Mireshghallah F, Uniyal A, Wang T, et al. Memorization in NLP fine-tuning methods. 2022. ArXiv:2205.12506
4 Charles Z, Ganesh A, McKenna R, et al. Fine-tuning large language models with user-level differential privacy. In: Proceedings of ICML 2024 Workshop on Theoretical Foundations of Foundation Models, 2024
5 Paik I, Choi J. The disharmony between bn and relu causes gradient explosion, but is offset by the correlation between activations. 2023. ArXiv:2304.11692