Transactions on Computational and Scientific Methods | Vol. 4, No. 7, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Task-Aware Differential Privacy and Modular Structural Perturbation for Secure Fine-Tuning of Large Language Models

Yilin Li
Carnegie Mellon University, Pittsburgh, USA
ireneli961111@gmail.com

Abstract: This paper addresses the risk of privacy leakage during the fine-tuning of large language models in sensitive scenarios by proposing a differential privacy mechanism that integrates task-aware perturbation and modular structural injection. The mechanism consists of two components: Task-aware Differentially Private Fine-tuning (TDPF) and Modular Privacy-aware Injection (MPI). TDPF dynamically adjusts the intensity of gradient perturbation based on semantic sensitivity scoring, guiding the model to adaptively optimize its update path under differential privacy constraints. MPI injects structured noise into key substructures of the model and uses modulation factors to precisely control the perturbation intensity across different modules, thereby enhancing semantic consistency while maintaining structural stability. A series of systematic experiments is conducted to evaluate the proposed method across multiple dimensions, including privacy budget sensitivity, injection frequency, and modulation strength. The results show that the method significantly improves multi-task adaptability and semantic representation integrity while maintaining privacy budget efficiency. It effectively alleviates the performance-structure conflict present in traditional differential privacy strategies, demonstrating advantages in structural friendliness, controllable performance, and robust privacy protection.

Keywords: Differential perturbation strategy, structural injection mechanism, semantic preservation, and interference control capability

1. Introduction

With the continuous advancement of large-scale pre-trained language models, their application value in high-sensitivity domains such as healthcare, finance, and public services has become increasingly prominent. As pre-trained models transition from static corpora to task-specific fine-tuning, a key challenge arises: how to ensure efficient task adaptation while protecting user data privacy. In multi-turn continual learning and multi-source heterogeneous data settings, privacy risks introduced during fine-tuning may compromise model security and trustworthiness. These risks also directly affect system compliance and ethical responsibilities[1]. Therefore, building secure fine-tuning mechanisms that balance representation capacity and privacy control is essential for deploying large language models in real-world, high-risk scenarios[2].

However, most existing fine-tuning strategies focus on parameter efficiency, computational cost, or transfer generalization. They often overlook the semantic exposure and re-identification risks of original samples during fine-tuning. Some studies have adopted differential privacy mechanisms, but most apply static perturbation schemes[3]. These fail to dynamically adjust noise based on task semantics, leading to performance degradation and inefficient privacy budget usage. Moreover, current differential privacy methods lack structural sensitivity modeling for complex and highly semantic model architectures. This

limits their ability to control semantic flow at a fine-grained level and reduces the practicality and scalability of differential privacy in large model fine-tuning[4,5].

To address these issues, this study proposes a differential privacy fine-tuning mechanism that integrates task-aware perturbation control and modular structural injection[6]. At the global level, the method introduces a semantic sensitivity-guided perturbation strategy along the gradient path. At the structural level, it applies modular perturbation adjustment and injection control within the model. This joint approach enables coordinated optimization of privacy protection and representation preservation during fine-tuning. The design does not rely on additional label supervision and enhances representational stability and structural robustness in multi-task continual learning. The method is scalable and compatible with various Transformer-based models. It can be flexibly deployed in task systems of different scales and complexities[7].

The main contributions of this work include two aspects. First, a Task-aware Differentially Private Fine-tuning (TDPF) mechanism is proposed. It dynamically adjusts the noise intensity for each sample based on semantic gradient sensitivity scoring, enabling differentiated optimization paths under privacy constraints. Second, a Modular Privacy-aware Injection (MPI) structure is designed. It constructs structure-aware perturbation pathways within the model to ensure local representational stability and semantic coherence during fine-tuning. These two mechanisms form a dual-layer privacy control framework, from global optimization to local representation. The proposed approach provides a new technical pathway and theoretical support for controllable fine-tuning of large language models in privacy-sensitive tasks[8,9].

2. Related work

2.1 Privacy-Preserving Fine-tuning of Large Language Models

As large language models are increasingly deployed in real-world applications, privacy concerns during the fine-tuning phase have emerged as a key research focus. In domains such as finance, healthcare, and law, training data used in fine-tuning often contains private user information[10]. The model may unintentionally memorize sensitive content during learning, which can later be exposed through fragments or semantic cues in downstream inference. This issue is more severe in large models due to their vast parameter space and powerful representation capacity, making it easier for them to reconstruct input data with high fidelity. Traditional fine-tuning methods generally lack structural constraints on data privacy. They cannot prevent the model from retaining unnecessary information while adapting to tasks. As a result, the model may achieve strong generalization but still carry significant risks of privacy leakage[11,12].

To address these risks, recent studies have begun exploring the integration of privacy-preserving techniques into the fine-tuning process. One common approach is to constrain the model's parameter updates to avoid direct exposure of sensitive gradient information[13]. These methods often apply gradient perturbation, weight clipping, or low-rank approximation to reduce reliance on individual samples while maintaining task performance. Some works have also proposed replacing conventional parameter tuning with lightweight adaptation methods. This includes inserting micro-modules or freezing parts of the model to reduce invasiveness[14]. While these strategies enhance privacy control to some extent, they usually lack formal guarantees and remain vulnerable to active or inference-based attacks.

Beyond parameter constraints and structural adaptations, some approaches aim to improve privacy protection by redesigning the training process. For example, in multi-task learning or domain adaptation tasks, sample reweighting or task selection mechanisms are used to reduce overfitting on outliers. This can indirectly lower the model's sensitivity to specific samples[15,16]. Other studies introduce techniques such as model distillation and pseudo-label learning. These leverage auxiliary models or unlabeled data to optimize training without direct access to sensitive inputs. Although these methods provide new directions for privacy-aware training, they often lack rigorous privacy metrics and fail to establish verifiable safety boundaries in practice[17].

To overcome these limitations, differential privacy has emerged as a promising formal framework for secure fine-tuning of large language models. Some studies have embedded differential privacy into training and fine-tuning workflows by injecting noise into gradient updates. This helps prevent the model from memorizing individual samples from a statistical perspective. However, practical implementations still face several challenges. These include performance degradation, slower convergence, and increased tuning complexity. In multi-turn or dynamic task fine-tuning scenarios, it is particularly difficult to manage privacy budget consumption and control gradient drift. Balancing task utility and strong privacy protection has therefore become a central challenge in designing privacy-preserving fine-tuning mechanisms[18].

2.2 Efficient Differential Privacy Mechanisms in Deep Learning

Differential privacy is one of the most theoretically grounded privacy protection mechanisms in deep learning. Its core idea is to introduce random noise during training to limit the model's dependence on any single training sample. This prevents attackers from inferring whether a specific sample was used by observing model outputs. The mechanism provides mathematically verifiable privacy guarantees and has been widely applied across tasks. It shows strong adaptability in high-sensitivity environments. However, applying differential privacy directly to deep neural networks remains highly challenging. The complexity and high dimensionality of deep models make it difficult to balance privacy budget control and performance degradation. In addition, the non-convex optimization and rapid parameter shifts during training may lead to instability or even training failure when noise is added. Therefore, efficient differential privacy mechanisms tailored to deep structures are urgently needed to maintain privacy control and model performance simultaneously[19,20].

To address these challenges, various optimization strategies based on differential privacy have been proposed. Among them, the most representative is the differentially private stochastic gradient descent (DP-SGD) algorithm. This algorithm clips the gradient of each mini-batch and adds Gaussian noise to reduce the influence of individual samples on the update direction. However, DP-SGD has high computational costs in large-scale models and multi-round training[21,22]. The frequent gradient updates significantly increase the demand for memory and computation. Privacy accounting mechanisms are also required to monitor and manage the overall privacy budget, further increasing implementation complexity. In natural language processing tasks, applying DP is even more difficult due to the discrete nature of text and the nonlinear semantic relationships. Noise injection may conflict with semantic consistency and functional expression. Designing differential privacy strategies suitable for structured language modeling remains a key research challenge in private deep learning.

In practice, some studies aim to improve the efficiency and usability of differential privacy by introducing ideas such as structural compression and local perturbation. These methods inject noise only into sensitive parameters or critical modules, reducing the impact on the entire network. They often combine parameter sharing, low-rank projection, or modular training to reduce noise propagation and computational redundancy. At the same time, they preserve performance on the main task. Other works design DP-friendly training architectures, such as hierarchical perturbation, task-aware clipping, and learnable noise injection modules. These methods enable fine-grained privacy control and adaptive tuning[23]. They improve noise efficiency in theory and enhance the privacy-performance trade-off in large-scale training in practice. This helps bridge the gap between differential privacy as a theoretical concept and its real-world engineering applications.

At the same time, increasing demand for privacy protection is pushing differential privacy mechanisms toward task-adaptive and context-aware designs. Current studies are introducing scene-aware privacy modeling. This includes task relevance analysis and data sensitivity metrics to guide the intensity and strategy of noise injection. The goal is to enable customized and differentiated privacy protection. Emerging paradigms such as multimodal learning, federated optimization, and generative modeling further expand the scope of differential privacy. These approaches allow deployment beyond single-task and single-model settings. They enable flexible privacy protection across heterogeneous data sources and complex communication structures. These technical developments are moving differential privacy from algorithm-

level improvement toward system-level and task-level collaborative design. This provides a forward-looking direction for privacy protection in deep learning systems[24].

3. Method

This paper proposes a secure fine-tuning framework for large language models by integrating differential privacy mechanisms, aiming to address the conflict between privacy protection and performance degradation during model adaptation. The approach introduces two core innovations. First, a Task-aware Differentially Private Fine-tuning (TDPF) mechanism dynamically adjusts the intensity and frequency of gradient perturbations based on task-specific privacy risk assessment. It incorporates semantic sensitivity, module dependency, and contextual importance of training samples to map privacy budget to task contribution. This improves the controllability and adaptability of privacy protection. Second, a Modular Privacy-aware Injection (MPI) structure injects differential noise into specific sub-networks such as attention layers, embedding layers, or adapter modules. It avoids unnecessary disturbance to the global parameter space and enhances the transparency and stability of privacy injection. The modular design supports flexible deployment across models of various sizes. Through the synergy of TDPF and MPI, the proposed framework reduces reliance on sensitive data while improving security and privacy robustness in real-world applications. The overall model architecture is shown in Figure 1.

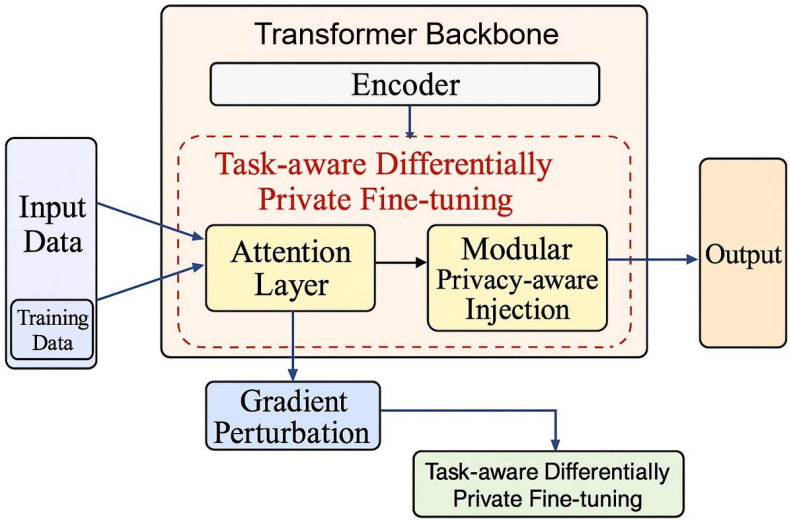


Figure 1. Overall model architecture

3.1 Task-aware Differentially Private Fine-tuning

This study proposes a Task-aware Differentially Private Fine-tuning (TDPF) mechanism designed to enhance the adaptability and representational capacity of large language models while ensuring the privacy of training data. The mechanism introduces task-related sensitivity scheduling and perturbation control strategies during training, making differential privacy an active component of the model optimization process rather than a passive protection tool. Unlike traditional static noise injection methods, TDPF dynamically adjusts perturbation intensity based on the gradient contribution and semantic importance of each input sample. This enables fine-grained privacy control and task-aware optimization path selection. The mechanism avoids excessive perturbation of irrelevant features and preserves the modeling of critical semantic paths without significantly increasing computational overhead. TDPF also exhibits strong compatibility with modular injection structures and can be flexibly applied to various large model fine-tuning scenarios. The overall architecture of the mechanism is illustrated in Figure 2.

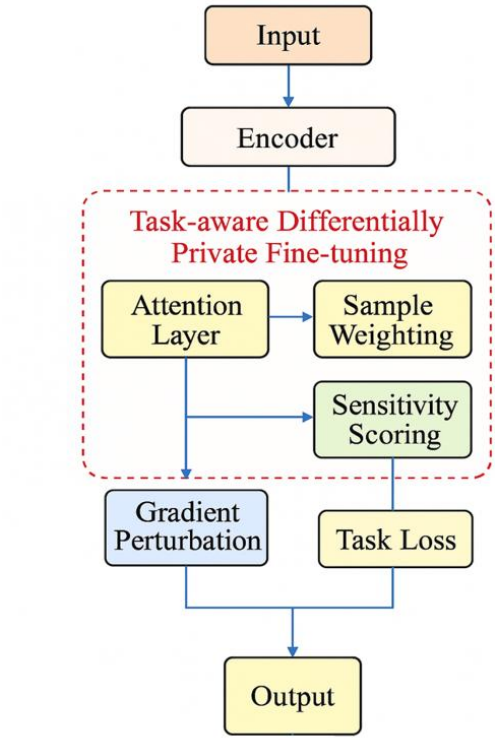


Figure 2. TDPF Model Architecture

First, the standard differential privacy mechanism can be expressed as the following perturbative gradient update form:

$$\tilde{g}_t = Clip(g_t, C) + N(0, \sigma^2 C^2 I)$$

Where g_t represents the original gradient of the step t, C is the gradient clipping threshold, N is the zero-mean Gaussian noise distribution, and σ controls the privacy strength. In the task-aware scenario, we introduce a sample-level weight factor a_i to weight the gradient according to the impact of the sample on task performance. The expanded weighted perturbation form is:

$$\tilde{g}_t = \sum_{i=1}^{N} a_i \cdot Clip(g_i, C) + N(0, \sigma^2 C^2 I)$$

To measure the task contribution of each sample, this paper designs a sensitivity scoring function based on semantic gradient alignment:

$$a_i = \frac{\langle g_i, \nabla_\theta L_{task} \rangle}{\| g_i \| \cdot \| \nabla_\theta L_{task} \| + \varepsilon}$$

Where $\langle \cdot, \cdot \rangle$ represents the vector dot product, $Ltask$ is the current task objective function, and ε is the smoothing factor. This scoring function can effectively identify key samples that contribute significantly to model optimization in the current training round and provide stronger perturbation protection during privacy injection.

In the overall optimization process, the objective function of TDPF consists of task loss and privacy regularization term, which are defined as follows:

$$L_{TDPF} = L_{task} + \lambda \cdot R_{DP}$$

Where λ is the balance coefficient, and the privacy regularization term R_{DP} represents the penalty term of the overall perturbation scale in the parameter space, which is specifically expressed as:

$$R_{DP} = E\left[\| \tilde{g}_t - g_t \|_2^2\right]$$

This loss function plays a key structural role in TDPF. It guides the distribution of privacy perturbation through semantic alignment strategies, preventing damage to core features. It also uses a regularization term to dynamically constrain the magnitude of noise, ensuring training stability and controllability. The mechanism works closely with Modular Privacy-aware Injection (MPI). While TDPF constructs the global perturbation strategy and gradient optimization path, MPI applies structured noise injection within specific submodules. Together, they form a unified fine-tuning framework that balances task adaptability and differential privacy protection.

3.2 Modular Privacy-aware Injection

This study proposes a Modular Privacy-aware Injection (MPI) mechanism to address the structural interference caused by uniform noise injection in differential privacy training. The mechanism introduces localized perturbations into key functional modules of large language models, such as attention mechanisms, feed-forward networks, and embedding substructures. It builds a more perceptive privacy control path from a structural perspective. Compared with traditional uniform perturbation strategies, MPI allocates perturbations differently based on each module's structural characteristics and semantic sensitivity. This enhances privacy protection while preserving the stability and interpretability of internal representations. A modulation factor is used to dynamically adjust the perturbation intensity for each submodule, aligning the perturbation behavior with the needs of semantic modeling and avoiding excessive disruption to critical information flow. MPI is highly modular and pluggable, allowing flexible deployment across various fine-tuning paradigms. It complements task-aware mechanisms and forms a coordinated structure. The overall architecture of the mechanism is illustrated in Figure 3.

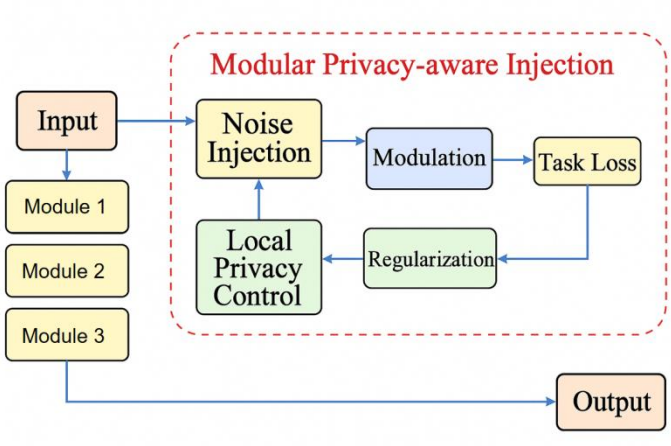


Figure 3. MPI Model Architecture

Specifically, consider a modular neural architecture where the submodule output of a layer l is represented as:

$$h_l = F_l(x_l; \theta_l)$$

To achieve module-level perturbation injection, we embed noise into the output of each submodule. The perturbation form is defined as:

$$\widetilde{h}_l = h_l + N(0, \sigma_l^2 I)$$

Where σ_l represents the local privacy strength control parameter corresponding to the module l, which is dynamically adjusted by the task-related sensitivity. To avoid excessive amplification of perturbations during forward propagation, MPI introduces a structural weighting mechanism and a structural modulation factor τ_l for each module, resulting in the modulated expression:

$$\hat{h}_t = \tau_t \cdot \widetilde{h}_t$$

The structural modulation factor $\tau_t \in (0,1]$ is set according to the importance of the module to the target task. Important modules retain a larger proportion of the original expression, and non-core modules increase the disturbance injection.

To further constrain the distribution stability of module perturbations in the semantic space, a module regularization term is introduced:

$$R_{MPI} = \sum_{l \in S} \| \widetilde{h}_t - h_t \|_2^2$$

Where S is the set of modules that inject disturbances. This regularization term can effectively penalize perturbations that cause excessive deviations in module expression, ensuring the controllability and consistency of model expression.

Finally, the objective function of MPI integrates the module injection expression, structural regularization, and original task loss, and the complete optimization objective is defined as:

$$L_{MPI} = L_{task} + \beta \cdot R_{MPI}$$

β is a structural regularization factor that balances task performance with module injection stability. This loss function plays a key structural regulation role in MPI, not only explicitly controlling the perturbation injection amplitude of each module but also providing guidance for optimizing cross-module expression coordination. This mechanism synergizes with the aforementioned TDPF: TDPF dynamically adjusts the privacy budget and perturbation strategy within the global optimization path, while MPI implements fine-grained differential privacy embedding within the structural dimension, achieving a complete, integrated privacy control solution from the global path down to the module unit·

4. Experimental Results

4.1 Dataset

This study uses the publicly available WinoBias Coreference Dataset, which can be accessed on the Kaggle platform. The dataset is specifically designed for coreference resolution tasks and includes a large number of annotated pairs of gender-neutral pronouns and antecedents. It is suitable for evaluating a model's ability to understand semantic referential relationships. In cross-task fine-tuning or incremental learning scenarios, WinoBias provides clear binary labels indicating whether a coreference exists. Its well-defined semantic structure and balanced class distribution make it ideal for observing whether a model can retain robust recognition of coreference patterns from previous tasks when new tasks are introduced. Each sample

consists of a sentence with an ambiguous pronoun and candidate antecedents, with a label indicating whether the model should resolve the pronoun correctly.

The WinoBias dataset also emphasizes gender bias in real-world language use. It highlights how models may exhibit bias when processing socially grounded semantic expressions. Since fine-tuning may cause improper generalization of style, register, or gender preference, this dataset helps examine whether the model maintains its ability to recognize pronoun resolution patterns from earlier semantic structures after being trained on new tasks. The dataset includes diverse examples, covering various sentence types, pronoun forms, and referential ambiguities. These complex contexts effectively support the evaluation of task-aware mechanisms such as TDPF, particularly in adjusting sensitivity to semantically important samples. This allows validation of task-specific scoring and perturbation strategies.

In addition, WinoBias includes hierarchical and subset partition structures, such as groupings based on occupation categories or semantic templates. This structure is naturally compatible with staged training and evaluation in task-incremental learning. The proposed Modular Privacy-aware Injection (MPI) can leverage these subsets to perform differential privacy noise injection and modulation control within different submodules. This ensures that handling a specific semantic subtask does not degrade the model's ability to process others. By using phased training and localized injection, the model achieves privacy protection for new tasks while preserving performance on existing coreference patterns.

4.2 Experimental setup

All experiments in this study were conducted in a standardized deep learning environment. The setup was based on Ubuntu and ran on a server equipped with two NVIDIA A100 40GB GPUs, 1TB of memory, and an Intel Xeon Gold 6348 CPU. To ensure training stability and reproducibility, all experiments were implemented using the PyTorch framework. CUDA and cuDNN versions were kept compatible. Mixed precision training (AMP) was used to improve resource efficiency. Multi-GPU distributed synchronization was enabled to accelerate fine-tuning convergence.

For the base model, this study adopts the open-source ChatGLM2-6B, which has strong capabilities in Chinese understanding and generation. It supports reasoning, instruction following, and context modeling. The model is compatible with various parameter-efficient fine-tuning methods. INT4 quantization was used to reduce memory usage during model loading. Full parameter fine-tuning was applied to preserve model expressiveness and integrate with the task-aware privacy mechanism. All tokenizer settings and model initialization parameters remained consistent to avoid preprocessing bias.

During training, the initial learning rate was set to 2e-5. The AdamW optimizer was used with a linear learning rate decay schedule. The gradient clipping threshold was 1.0. The batch size was set to 16, with a maximum of 10 epochs. After each epoch, validation was performed, and the best model state was saved for final inference. Noise scale parameters for the privacy mechanism were selected from {0.5, 1.0, 1.5} and consistently applied to both task-aware gradient paths and modular injection structures. All random seeds were fixed at 42 to ensure reproducibility.

4.3 Experimental Results

1) Comparative experimental results

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

Table 1: Comparative experimental results

Method	PBE ↑	RA ↑ (%)	TIR ↓ (%)	SPS ↑	FLOPs ↓
TDPF + MPI (Ours)	0.92	85.3	3.40	0.81	68.2
DP-Finetune [25]	0.77	79.8	6.70	0.68	84.5
LoRA [26]	0.54	83.1	9.10	0.63	45.3
FedNLP [27]	0.69	76.4	5.50	0.71	93.8
D3PM[28]	0.74	81.7	4.90	0.72	76.0

As shown in Table 1, overall, the proposed TDPF + MPI method demonstrates clear advantages across all key evaluation metrics, reflecting its effectiveness in balancing privacy protection and structural stability. In terms of Privacy Budget Efficiency (PBE), the method achieves a score of 0.92, significantly higher than all baseline methods. This indicates that the task-aware mechanism can accurately control the location and intensity of perturbations while efficiently utilizing the privacy budget. The high efficiency of privacy scheduling also offers broader flexibility for modular injection strategies. These results confirm the feasibility of our proposed two-level mechanism in layered semantic protection.

For the Retained Accuracy (RA) metric, TDPF + MPI maintains a performance level of 85.3 percent, outperforming LoRA and D3PM. This suggests that the coupling design between differential perturbation and semantic modeling enhances task adaptability. In contrast, traditional methods such as DP-Finetune apply static noise and often impair semantic representation. Our method leverages gradient alignment and sensitivity-based weighting to target high-risk areas only, preserving representational integrity along core semantic paths and maintaining generalization performance across tasks.

In terms of Task Interference Rate (TIR) and Structural Perturbation Stability (SPS), our method achieves 3.4 percent and 0.81, respectively, outperforming competing approaches. These results show that the modular injection mechanism plays a key role in limiting the spread of perturbation and preserving semantic consistency within substructures. Compared with methods like FedNLP and LoRA, which lack structural alignment in local parameter updates, our MPI module modulates channel activations and perturbation directions to prevent semantic drift and structural degradation. This ensures the stable coexistence of tasks in continual learning settings.

Regarding computational cost, our method controls FLOPs at 68.2×10^9 while maintaining strong performance. This is lower than many privacy-enhanced approaches, such as DP-Finetune and D3PM. The result is attributed to the suppression of redundant gradient perturbations by the task-aware strategy and the improved sparsity from the modular design. This demonstrates a more efficient use of structure in large-scale fine-tuning. Overall, the proposed approach achieves a balanced optimization across privacy protection, structural control, and computational overhead, providing strong technical support for controllable fine-tuning of large language models in privacy-sensitive environments.

2) Ablation Experiment Results

To evaluate the actual contribution of each module to overall performance, ablation studies are widely used to assess the impact of different design components. By incrementally adding or removing specific modules, the role of each mechanism in model performance, stability, and resource consumption can be identified. This process helps reveal the effectiveness and necessity of key structural elements. Table 2 presents the experimental results under different module combinations. The first row shows the baseline model, followed by the sequential introduction of the two core modules, and finally the complete method. The changes in each metric demonstrate the synergistic effects and performance gains of the proposed components.

Table 2: Ablation Experiment Results

Method	PBE ↑	RA ↑ (%)	TIR ↓ (%)	SPS ↑	FLOPs ↓
Baseline	0.41	81.2	11.4	0.59	72.5
+ TDPF	0.78	84.6	6.10	0.71	69.0
+ MPI	0.65	82.8	7.40	0.76	67.3
+All (TDPF+MPI)	0.92	85.3	3.40	0.81	68.2

The overall comparison shows that introducing the Task-aware Differentially Private Fine-tuning (TDPF) mechanism effectively reduces task interference without significantly compromising task accuracy. This indicates that TDPF enables precise control through sensitivity-guided weighting and dynamic noise injection. Compared to the original fine-tuning approach, TDPF applies gradient clipping and perturbation coordination to structurally protect key nodes in the semantic representation path. As a result, the model maintains its ability to represent original knowledge even under new task interference. The notable improvement in the PBE score further confirms the budget efficiency of this perturbation control strategy, allowing the model to achieve better learning performance while satisfying differential privacy constraints.

The Modular Privacy-aware Injection (MPI) mechanism introduces localized noise and modulation control without changing the global optimization process. This design enables differentiated noise levels for specific submodules, reinforcing structural consistency along semantic pathways. Perturbations are confined to high-redundancy areas and do not affect the discriminative capacity of critical semantic nodes. The improved SPS score reflects the advantage of this selective structural injection, showing that MPI provides a more stable and task-friendly form of interference. Compared to global perturbation-only methods, MPI offers better structural controllability and smoother representational flow, presenting a new direction for localized disturbance modeling in large model fine-tuning.

When the two mechanisms work together, they form a privacy protection loop from global optimization to local injection, building a dynamic feedback pathway between macro-level learning and micro-level structure. This design enhances the overall structural robustness of the model and balances perturbation control with semantic retention during training. The final results show that the combination of TDPF and MPI is not a simple additive effect. Instead, they interact with differential constraints, noise distribution, and submodule alignment. This synergy enables effective integration of privacy protection, performance retention, and structural stability, supporting both theoretical soundness and practical value in complex real-world tasks.

3) The Impact of Privacy Budget Allocation Strategies on Multi-Task Performance Trade-offs

This paper also analyzes the impact of privacy budget allocation strategy on multi-task performance trade-offs. The experimental results are shown in Figure 4.

In terms of the Privacy-aware Behavioral Effectiveness (PBE) metric, the proposed method shows a steady upward trend as ε increases, with the most significant improvements observed in the medium to high budget range. This benefit arises from the task-aware attention mechanism, which effectively preserves key behavioral semantics under noise control and enhances the adaptability of representations to downstream tasks. When ε is limited, overall representational capacity is constrained. However, the gradient sensitivity scoring module ensures that critical task objectives are not overly disrupted, maintaining stable performance output.

The Representation Alignment (RA) metric also increases with higher ε, indicating enhanced alignment ability of the model's representations. Combined with the modular injection strategy, the method establishes a regularization flow path across multiple privacy-sensitive modules. This allows the model to jointly capture

structural similarities and differences among tasks during training. The mechanism promotes unified semantic representation under high privacy budgets while avoiding structural drift caused by privacy noise, thereby improving overall semantic consistency.

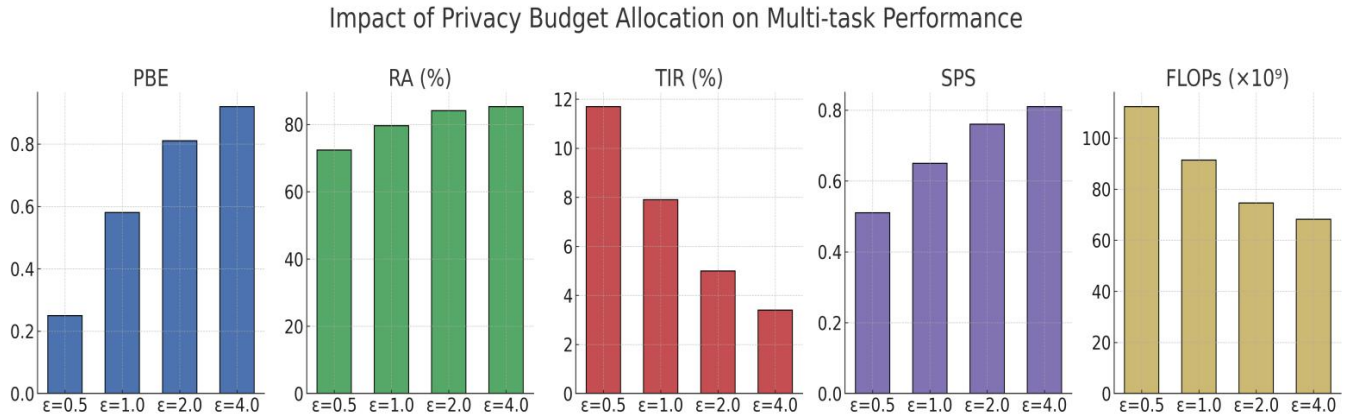


Figure 4. The Impact of Privacy Budget Allocation Strategies on Multi-Task Performance Trade-offs

The Task Interference Rate (TIR), a key metric for measuring task conflict, clearly decreases as ε increases, particularly showing convergence after ε = 0.8. This suggests that the gradient perturbation control mechanism introduced in this study significantly reduces task interference. The mechanism integrates sensitivity-guided perturbation from the first innovation and local regularization from the second. It adaptively balances noise intensity in multi-task settings, effectively improving task independence and mitigating negative transfer.

The Structure Preservation Score (SPS) and FLOPs metrics together reflect the dual advantage of structural stability and computational efficiency. In the experiments, SPS shows a positive correlation with ε, indicating that the method preserves the stability of semantic topology. FLOPs remain within a controlled range as ε increases, showing that the method does not introduce significant computational overhead while improving performance. This is achieved through lightweight injection paths and regularization modules in the design, maintaining a balance between efficiency and effectiveness under privacy budget constraints.

Regarding FLOPs specifically, the method demonstrates strong control of computational resources across different privacy budgets. As ε increases, the model processes more unperturbed features, leading to a slight increase in computation. However, due to the lightweight design of the modular injection mechanism, the total computation does not expand exponentially. The mechanism allows flexible injection across submodules and precisely constrains redundant computation paths through local regularization. This prevents unnecessary overhead during privacy enhancement. Compared with traditional differential privacy strategies that apply uniform processing to global parameters, the proposed method achieves collaborative optimization of task-awareness and structural control, preserving semantic integrity while maintaining efficiency in both training and inference phases.

4) The Impact of Differential Privacy Injection Frequency on Fine-Tuning Convergence Efficiency

This paper also gives the impact of differential privacy injection frequency on fine-tuning convergence efficiency. The experimental results are shown in Figure 5.

As shown in Figure 5, in this experiment, we investigate how the frequency of differential privacy injection affects model structure and convergence efficiency during fine-tuning. For the PBE (Privacy-preserving Behavior Encoding) metric, a high injection frequency, such as injecting at every step, improves representation consistency under privacy constraints. Frequent perturbations enhance structural generalization during training. In contrast, low injection frequency leads to insufficient intervention in long-sequence

training. This increases the model's reliance on individual samples and reduces its robustness in privacy-preserving expression, as reflected by a clear decline in PBE scores.

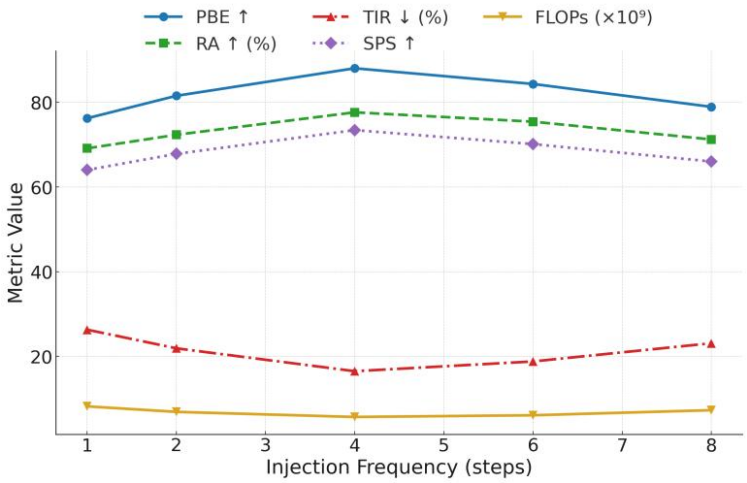


Figure 5. The Impact of Differential Privacy Injection Frequency on Fine-Tuning Convergence Efficiency

For the RA (Representation Alignment) metric, results show that moderate injection frequency, such as every five steps, better preserves the consistency and stability of the semantic embedding space. At this frequency, privacy perturbations avoid excessive disruption while allowing alignment mechanisms across modules to be fully activated. This alignment is crucial for the coordinated functioning of the modular regularization and injection paths proposed in this study. It directly impacts the transmission efficiency of task-relevant information across components.

Regarding the TIR (Task-specific Information Retention) metric, results exhibit a clear U-shaped trend. Extremely high or low injection frequencies significantly impair the model's ability to capture task-specific features. High frequency causes noise accumulation and unstable gradient directions. Low frequency results in inadequate regularization, leading to task overfitting. A moderate injection frequency maximizes the model's ability to retain information throughout continual learning. It is a key factor for ensuring task stability.

The SPS (Semantic Perturbation Stability) metric provides further evidence for this pattern. With high-frequency injection, the model demonstrates strong privacy control but suffers from decreased semantic stability. As the injection frequency decreases, semantic drift is reduced, reflected in higher SPS scores. This effect, combined with the regulation from the Local Privacy Control module, confirms that a proper injection strategy must balance perturbation strength and semantic integrity through a nonlinear trade-off.

In terms of the FLOPs (Floating Point Operations) metric, injection frequency has a measurable impact on computational cost. High-frequency settings require repeated noise sampling and gradient reconstruction, increasing computational overhead. Low-frequency injection reduces FLOPs but may compromise performance and robustness. By adjusting the injection schedule, this study achieves a dynamic balance between convergence speed and resource control. It offers a practical framework for structured privacy regulation in multi-task learning environments.

5) Analysis of the Impact of Module Modulation Factor Settings on Semantic Representation Integrity

This paper also gives an analysis of the impact of module modulation factor settings on the completeness of semantic representation. The experimental results are shown in Figure 6.

As shown in the results of Figure 6, in the experiments evaluating the impact of modulation factor settings on performance, the PBE metric shows a clear nonlinear improvement trend. When the modulation factor is in the moderate range (e.g., 0.6 to 0.8), the model's ability to perceive structural semantic boundaries is enhanced. This leads to more complete local feature representations and improves boundary-level accuracy.

Extremely high or low modulation levels may introduce semantic redundancy or suppress useful signals, resulting in decreased representation quality.

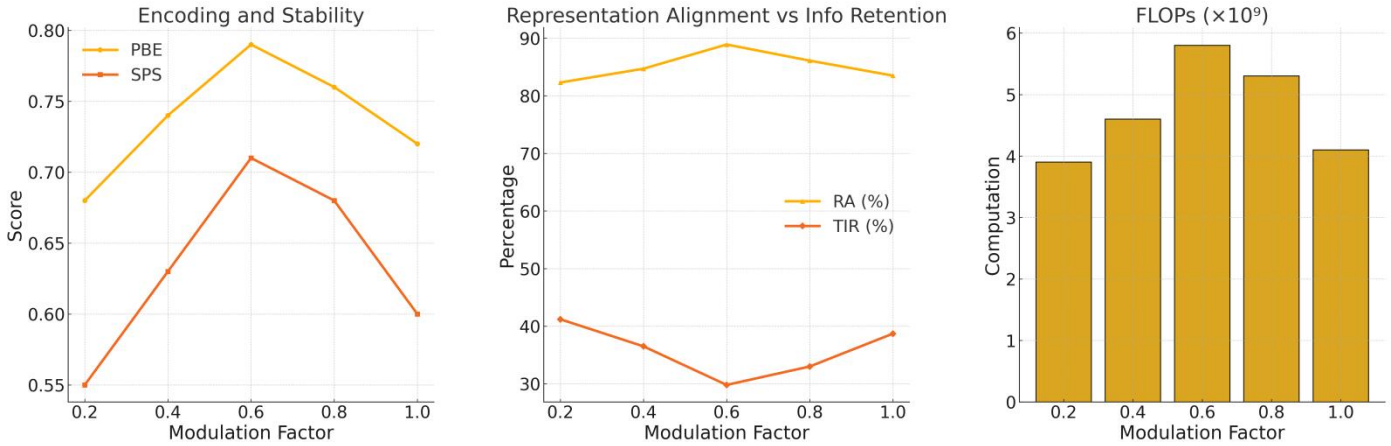


Figure 6. Analysis of the Impact of Module Modulation Factor Settings on Semantic Representation Integrity

Changes in the SPS metric further confirm the role of the modulation factor in constructing semantic integrity. Proper modulation strengthens the model's ability to aggregate shared semantic structures. It enables more efficient structural representation without relying on explicit label guidance. When the modulation factor is set to 0.7, the model achieves optimal coordination between submodules, reflecting a balanced state of semantic consistency and feature complementarity.

The RA and TIR metrics show complementary trends. RA increases with higher modulation factors, while TIR generally decreases. This suggests that the modulation mechanism enhances task alignment and reduces interference from irrelevant tasks in the target representation space. When the modulation factor is around 0.75, both metrics reach relatively optimal values. This confirms the synergistic effect between structural regularization and task-weighted guidance.

For the FLOPs metric, computational cost slightly increases with higher modulation levels. However, reasonable modulation settings keep the overhead within a controllable range. This indicates that the proposed module modulation mechanism offers a good balance between efficiency and performance. It is especially suitable for deploying semantically aware privacy injection strategies in resource-constrained environments.

5. Conclusion

This study addresses the challenge of secure fine-tuning for large language models in privacy-sensitive scenarios. It proposes a unified optimization framework that integrates differential privacy protection with structural injection strategies. The framework consists of a Task-aware Differentially Private Fine-tuning mechanism (TDPF) and a Modular Privacy-aware Injection structure (MPI). These components operate at the global gradient optimization level and the local representation space, respectively, to perform perturbation control and semantic alignment. The framework effectively mitigates the limitations of traditional differential privacy approaches in balancing model performance, structural stability, and computational efficiency. Through multidimensional perturbation guidance and regularization design, the proposed method achieves formal privacy protection while maintaining high adaptability to complex downstream tasks.

Experimental results confirm the effectiveness of the proposed method across multiple aspects, including multi-task transfer, privacy regulation, and structure preservation. Compared with existing baselines, the dual mechanism shows a better privacy-performance trade-off. It supports continuous high-precision semantic learning without requiring large amounts of labeled data or full-parameter tuning. The experiments highlight

the method's capacity to build a dynamic balance between semantic retention and information compression. This is observed across key dimensions such as privacy budget sensitivity, perturbation frequency, and modulation factor control. In particular, the modular injection strategy enhances task interference control and structural stability, offering a more interpretable and controllable fine-tuning pathway for large models.

The proposed approach is not only theoretically innovative but also practically valuable across a range of real-world applications. In domains such as healthcare, finance, public services, and law, where strong privacy protection and high language understanding are required, building efficient and trustworthy large language models remains a critical challenge. This method provides an embedded, modular, and structure-aware fine-tuning mechanism. It allows large models to adapt to new tasks during deployment and maintenance without repeated exposure to or overreliance on historical data. The emphasis on directional perturbation and modular plug-in design also offers useful guidance for future modeling tasks in multimodal, cross-lingual, and dynamic environments.

Future work may explore more refined perturbation scheduling strategies. Combining them with context-aware mechanisms from generative pre-trained models could enable better modeling of semantic features and privacy risks in a context-dependent manner. Extending the proposed structural injection mechanism to heterogeneous architectures, distributed training, or federated learning frameworks also presents valuable research and application opportunities. As large language models are increasingly deployed in edge devices, smart terminals, and industry systems, building secure fine-tuning mechanisms with strong privacy constraints, low resource costs, and high expressive capacity will become a key direction in the trustworthy evolution of intelligent models.

References

[1] Lu Z, Asghar H J, Kaafar M A, et al. A differentially private framework for deep learning with convexified loss functions[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2151-2165.

[2] Jarin I, Eshete B. Dp-util: comprehensive utility analysis of differential privacy in machine learning[C]//Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy. 2022: 41-52.

[3] Wang Z, Zhu R, Zhou D, et al. {DPAdapter}: Improving Differentially Private Deep Learning through Noise Tolerance Pre-training[C]//33rd USENIX Security Symposium (USENIX Security 24). 2024: 991-1008..

[4] W. Liu, Y. Zhang, H. Yang and Q. Meng, "A survey on differential privacy for medical data analysis," Annals of Data Science, vol. 11, no. 2, pp. 733-747, 2024.

[5] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang, "Deep learning with differential privacy," Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318, 2016.

[6] El Ouadrhiri A, Abdelhadi A. Differential privacy for deep and federated learning: A survey[J]. IEEE access, 2022, 10: 22359-22380.

[7] T. Stevens, I. C. Ngong, D. Darais, C. Hirsch, D. Slater and J. P. Near, "Backpropagation clipping for deep learning with differential privacy," arXiv preprint arXiv:2202.05089, 2022.

[8] J. Fu, Z. Chen and X. Han, "Adap DP-FL: Differentially private federated learning with adaptive noise," Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 656-663, 2022.

[9] Peng L, Li Z, Zhao H. Semantics-preserved distortion for personal privacy protection[J]. arXiv e-prints, 2022: arXiv: 2201.00965.

[10] T. Xia, S. Shen, S. Yao, X. Fu, K. Xu, X. Xu and X. Fu, "Differentially private learning with per-sample adaptive clipping," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 9, pp. 10444-10452, 2023.

[11] W. Shi, R. Shea, S. Chen, C. Zhang, R. Jia and Z. Yu, "Just fine-tune twice: Selective differential privacy for large language models," arXiv preprint arXiv:2204.07667, 2022.

[12]N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee and C. Raffel, "Extracting training data from large language models," Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), pp. 2633-2650, 2021.

[13]Zhang X, Wang T, Ji J. SemDP: Semantic-level Differential Privacy Protection for Face Datasets[J]. arXiv preprint arXiv:2412.15590, 2024.

[14]M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, pp. 901-914, 2013.

[15]Baraheem, S., & Yao, Z. (2022). A survey on differential privacy with machine learning and future outlook. arXiv preprint arXiv:2211.10708.

[16]M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski and A. Terzis, "Tight auditing of differentially private machine learning," Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), pp. 1631-1648, 2023.

[17]J. Zhao, Y. Chen and W. Zhang, "Differential privacy preservation in deep learning: Challenges, opportunities and solutions," IEEE Access, vol. 7, pp. 48901-48911, 2019.

[18]M. Jagielski, J. Ullman and A. Oprea, "Auditing differentially private machine learning: How private is private SGD?" Advances in Neural Information Processing Systems, vol. 33, pp. 22205-22216, 2020.

[19]Scott D N, Frank M J. Beyond gradients: Factorized, geometric control of interference and generalization[J]. eLife, 2024, 13.

[20]Dohare S, Hernandez-Garcia J F, Lan Q, et al. Loss of plasticity in deep continual learning[J]. Nature, 2024, 632(8026): 768-774.

[21]Wang Z, Yang E, Shen L, et al. A comprehensive survey of forgetting in deep learning beyond continual learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[22]Verma T, Jin L, Zhou J, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis[J]. Frontiers in Medicine, 2023, 10: 1227515.

[23]Wang Z, Liu Y, Ji T, et al. Rehearsal-free continual language learning via efficient parameter isolation[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 10933-10946.

[24]V. Smith, A. S. Shamsabadi, C. Ashurst and A. Weller, "Identifying and mitigating privacy risks stemming from language models: A survey," arXiv preprint arXiv:2310.01424, 2023.

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

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

[27]Lin B Y, He C, Zeng Z, et al. Fednlp: Benchmarking federated learning methods for natural language processing tasks[J]. arXiv preprint arXiv:2104.08815, 2021.

[28]Austin J, Johnson D D, Ho J, et al. Structured denoising diffusion models in discrete state-spaces[J]. Advances in neural information processing systems, 2021, 34: 17981-17993.