Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 4, 2024

Federated Distillation with Structural Perturbation for Robust Fine-Tuning of LLMs

Yujun Zou
University of California, Berkeley, Berkeley, USA
zouyujun526@gmail.com

Abstract: This paper proposes a federated fine-tuning framework that integrates differentiated distillation guidance and modular structure injection to address key challenges in distributed fine-tuning of large language models. These challenges include unstable semantic transfer, sensitivity to structural perturbations, and low communication efficiency. Without transmitting raw data, the framework introduces a differentiated distillation mechanism to guide local client models in aligning with the global semantic structure. This reduces representation drift under non-independent and identically distributed multi-task settings. Meanwhile, a modular structure injection mechanism applies structural perturbations to key components such as attention layers and feedforward networks. This guides the model to learn robust representations under local variation, enhancing the consistency and stability of cross-task representations. The two mechanisms are designed to be decoupled yet jointly optimized. They can be flexibly embedded into mainstream pre-trained language models and enable communication-efficient distributed knowledge optimization under the federated learning framework. Experiments on multiple task-incremental subsets verify the effectiveness of the proposed method. Through comprehensive main experiments, ablation studies, and hyperparameter sensitivity analyses, the model is evaluated across multiple dimensions, including semantic retention, structural stability, generalization ability, and parameter efficiency. The results show that the proposed method outperforms existing representative approaches and demonstrates strong practical value and adaptability.

Keywords: Federated distillation; structural perturbation; module injection; semantic alignment

1. Introduction

Driven by the rapid development of large language models, natural language processing systems are now capable of handling multi-task, multilingual, and cross-domain semantic understanding. As model sizes continue to grow, traditional centralized training and fine-tuning methods face increasing limitations in terms of resource cost, privacy protection, and scalability. In particular, domains such as healthcare, finance, and public services impose strict requirements on data security and on-device deployment[1]. How to achieve low-cost adaptation and efficient updating of large models has become a key barrier to the practical deployment of intelligent systems. Therefore, developing a distributed fine-tuning paradigm for large language models that supports privacy preservation, efficient collaboration, and structural adaptability holds significant practical value and research importance[2,3].

Although federated learning provides a foundational framework for collaborative modeling in distributed settings, most existing methods focus on lightweight architectures or small-scale models[4]. These approaches struggle to balance stability and efficiency when applied to large language models with high semantic complexity and structural heterogeneity. On one hand, unified distillation mechanisms often fail to accommodate semantic shifts across clients, resulting in weak knowledge transfer or misaligned learning. On the other hand, current fine-tuning strategies cannot model internal structural dynamics, leading to representational inconsistency or structural collapse during task increments or local perturbations. These limitations significantly affect the generalization and usability of large models in multi-task and cross-client scenarios[5,6].

To address these challenges, this paper proposes a distributed fine-tuning method for large language models that integrates federated distillation and structural injection. Within a federated optimization framework, the method introduces two components: a Differentiated Distillation Guidance (DDG) mechanism and a Modular Structure Injection (MSI) strategy. These components enhance semantic alignment and structural stability, respectively[8]. The method aims to solve key issues such as weak cross-client knowledge transfer and unstable representation spaces. It is modular, lightweight, and scalable. It can be applied to various language model architectures and supports secure and efficient task adaptation across diverse client environments[9].

This study presents two main innovations. First, the Differentiated Distillation Guidance (DDG) mechanism is proposed to guide client learning through dual pathways of semantic aggregation and semantic alignment. This enhances the stability of semantic transfer in non-independent and identically distributed scenarios. Second, the Modular Structure Injection (MSI) mechanism introduces structural perturbations into key sub-modules of the model. This guides the model to learn robust responses to local variations, improving consistency and robustness of the representation space during task switching and structural drift. Together, these two mechanisms provide structural support and optimization

pathways for enhancing the performance of large language models in distributed semantic modeling and incremental task fine-tuning[10,11].

2. Prior Work

2.1 Large Language Model Fine-tuning Algorithm

Federated learning is a distributed collaborative modeling paradigm designed to enable joint model training across devices or domains without transferring raw data. Its core idea is to update model parameters locally at each participant and then upload the updated parameters to a central server for aggregation[12]. This process builds a globally shared model. The mechanism offers natural advantages in protecting user privacy and reducing the risk of data leakage. It is particularly suitable for scenarios involving dispersed and sensitive data. As practical applications grow in complexity, the traditional federated averaging strategy has shown limitations when dealing with non-independent and identically distributed data and device heterogeneity[13]. This has driven researchers to explore more robust and scalable optimization mechanisms to enhance the applicability of federated learning in real-world environments.

Knowledge distillation was initially proposed as a model compression technique. It leverages soft label information generated by a high-capacity model to guide the training of a lightweight model. This allows for effective compression while maintaining performance. In federated learning, this idea has been re-integrated into a mechanism known as federated distillation. Instead of aggregating model parameters directly, federated distillation transmits soft prediction probabilities or intermediate representations[14,15]. This enables indirect knowledge sharing, significantly reducing communication overhead and improving adaptability to model heterogeneity. Especially in scenarios where client data is highly imbalanced, soft labels can effectively capture semantic commonalities across clients. This helps alleviate model performance fluctuations caused by sample distribution shifts.

The integration of distillation into the federated learning framework has led to various forms of knowledge exchange. Some methods rely on a central server to collect and distribute knowledge, while others adopt peer-to-peer distillation to support finer-grained knowledge sharing. In these approaches, clients do not expose their model parameters or gradients. Instead, they upload predictions generated from private data, promoting knowledge aggregation while preserving privacy. Moreover, as the distillation process can flexibly select different levels of features or representations for transmission, the mechanism is especially useful in bandwidth-constrained environments. It avoids the high cost associated with training large models and supports adaptation in heterogeneous systems with diverse client capabilities[16].

To improve the performance of federated distillation in complex settings, researchers have introduced designs such as multi-teacher, multi-stage, and cross-client distillation. These structures aim to enhance semantic alignment and model collaboration among clients. Based on this, variants of federated distillation have emerged, incorporating attention mechanisms, contrastive learning, and perturbation-based enhancement. These methods improve the accuracy and generalizability of knowledge transfer[17]. They also demonstrate scalability and stability in practical tasks, extending the use of federated distillation to natural language processing, image recognition, and speech modeling. Nevertheless, as models grow larger and tasks become more complex, challenges remain. Improving distillation efficiency, reducing communication costs, and enhancing model robustness are still key problems to be addressed[18].

2.2 Fine-Tuning of Large Language Models in Distributed Settings

Large language models have become fundamental components of modern artificial intelligence systems due to their strengths in semantic understanding, context modeling, and multi-task generalization. However, these models often contain billions or even hundreds of billions of parameters. The fine-tuning process consumes significant computational power, storage, and energy, far exceeding the capacity of ordinary computing devices. In resource-constrained or multi-terminal deployment scenarios, traditional full-parameter fine-tuning becomes overly cumbersome and cannot meet the demand for efficient and low-cost applications. At the same time, privacy concerns and data sovereignty issues have increasingly restricted centralized training on large-scale corpora. As a result, efficient fine-tuning of large language models in distributed environments has become a core challenge in practical applications[19,20].

To address these challenges, several parameter-efficient fine-tuning techniques have emerged. These include low-rank adaptation, prompt tuning, and modular injection. Such methods freeze the backbone parameters and only update a small number of sub-modules or soft prompts. They significantly reduce resource consumption during training. While maintaining model performance, they enhance the flexibility of the fine-tuning process[21]. However, when applied to distributed environments, these techniques face new problems. For example, under heterogeneous data and task settings, it remains difficult to ensure consistency across fine-tuned modules on different devices. Semantic drift during module updates can also occur. These issues directly affect model stability and generalization, limiting the applicability of lightweight tuning strategies in distributed settings[22].

In multi-terminal collaborative optimization, fine-tuning language models must also deal with device heterogeneity, limited communication bandwidth, and differences in client task objectives. Some studies have introduced periodic synchronization, knowledge aggregation, or incremental fusion on top of local fine-tuning. These approaches aim to balance global modeling capacity with local customization. However, they often require many communication rounds and frequent parameter synchronization, which increases the system burden. As model size continues to grow, the resource cost of a single synchronization round rises sharply[23,24]. This severely impacts overall training efficiency. Therefore, building a fine-tuning framework for large language models that is highly adaptive, communication-efficient, and stable in distributed environments has become an urgent need.

Given the high resource demands of fine-tuning large models and the constraints of distributed deployment, recent studies have turned attention to knowledge transfer and knowledge fusion in distributed optimization. By aligning local models with global knowledge, it is possible to improve model capacity without directly sharing parameters. In large language models, context-sensitive expression makes semantic consistency and representation compression key performance indicators. This drives researchers to explore structural awareness, semantic preservation, and perturbation control mechanisms for efficient fine-tuning in distributed scenarios. At the same time, enabling effective knowledge sharing and unified representation across clients has become a critical direction for improving model transferability and deployment feasibility[25].

3. Approach

This study proposes an efficient fine-tuning framework that integrates federated distillation with structure-aware strategies to address the challenges of high communication cost, parameter redundancy, and semantic misalignment in distributed fine-tuning of large language models. The framework includes two core innovations. First, a Differentiated Distillation Guidance (DDG) mechanism is designed to enhance the consistency and robustness of knowledge transfer by introducing cross-client semantic harmonization paths under privacy constraints. Second, a Modular Structure Injection (MSI) strategy is developed to apply structure-aware perturbations to key sub-modules of large language models, guiding the formation of generalizable semantic representations in multi-client environments. Together, these two modules alleviate representation drift in distributed optimization and provide structural support for building fine-tuning paradigms with low communication overhead and high efficiency. The overall model architecture is shown in Figure 1.

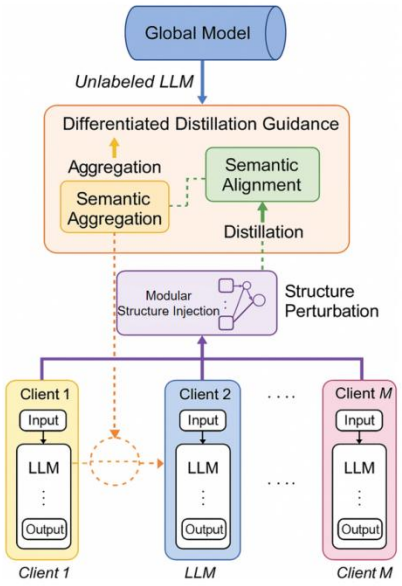

Figure 1. Overall model architecture

3.1 Differentiated Distillation Guidance

The Differentiated Distillation Guidance (DDG) mechanism proposed in this study aims to address common issues of semantic drift and representation inconsistency in distributed fine-tuning of language models. DDG builds a dual-path distillation framework of semantic aggregation and semantic alignment. Without transmitting raw data, it enables fine-grained soft knowledge interaction between client models and the global model. Specifically, the semantic aggregation path first generates a weighted summary of soft predictions from all clients to form a global semantic distribution that captures cross-domain commonalities. Then, the semantic alignment path applies joint constraints at both the probability and representation levels using the aggregated distribution and global intermediate features. This ensures that key semantic structures are preserved during local updates and that distributional shift is reduced. Compared to traditional single-path or unified distillation strategies, DDG dynamically perceives semantic differences among clients and adjusts distillation strength accordingly, enabling more robust knowledge transfer. This mechanism not only improves semantic consistency in multi-client collaborative training but also provides a stable representational foundation for subsequent modules such as structural injection and task increment. The full workflow and key components are shown in Figure 2.

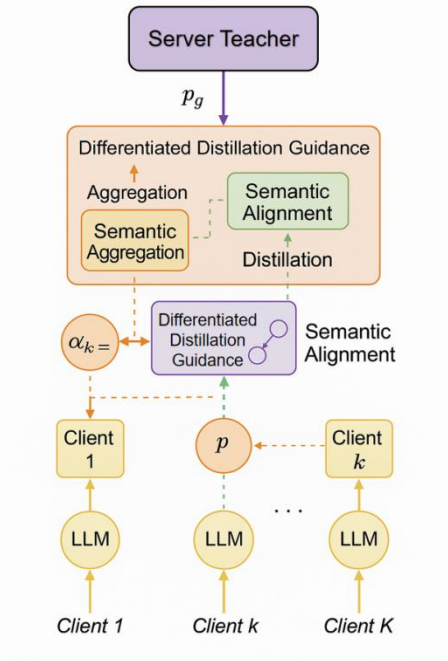

Figure 2. DDG Model Architecture

Let the client model be $f_k(x;\theta_k)$, where θ_k represents the model parameters of the k th client and x is the input sample. Each client represents the prediction of its local model for the input x as a probability distribution $p_k = f_k(x)$, and the corresponding global teacher model is $f_g(x;\theta_g)$, whose output is $p_g = f_g(x)$. To achieve cross-client distillation guidance, this study first constructs the distillation loss function as follows:

$$L_{distill} = \sum_{k=1}^{K} E_{x\sim D_k} KL(p_g(x) \| p_k(x))$$

Where $KL(\cdot\|\cdot)$ represents the Kullback-Leibler divergence, which measures the distance between the prediction distribution of the global model and the client model. This loss term ensures that the client model maintains consistency with the global semantics during local fine-tuning.

To further enhance the stability of semantic representation, semantic alignment loss is introduced to align representations by comparing the Euclidean distance between the intermediate layer representations h_g and h_k:

$$L_{align} = \sum_{k=1}^{K} E_{x\sim D_k} \| h_k(x) - h_g(x) \|_2^2$$

This loss term is designed to constrain the representation space learned by client models to remain structurally consistent with the global model. It reduces semantic drift caused by differences in data distribution.

Taking into account the heterogeneity between client models, DDG further introduces a differentiated distillation weight function to dynamically adjust the distillation strength according to the degree of semantic deviation of each client:

$$a_k = \frac{1}{1+\exp(-\tau \cdot \delta_k)} \quad \text{with} \quad \delta_k = \| p_k(x) - p_g(x) \|_2$$

Where δ_k represents the semantic deviation between the client and the global model, γ is the adjustment parameter, and $a_k \in (0,1)$ represents the client distillation guidance strength. This weighting factor adaptively adjusts the influence range of the distillation target, achieving personalized control during the knowledge transfer process.

During the distillation guidance process, it is also necessary to fuse the soft predictions from all clients for semantic aggregation to construct an aggregated "teacher" representation:

$$\bar{p}(x) = \sum_{k=1}^{K} w_k \cdot p_k(x), with \sum_{k=1}^{K} w_k = 1$$

w_k can be allocated based on historical performance, data volume, or number of training rounds to balance client contributions. This aggregate distribution is used to assist global model training and maintain its ability to perceive multi-source semantics.

Finally, the overall optimization objective of the DDG module is composed of the weighted loss functions above to form a complete guided objective function:

$$L_{DDG} = \lambda_1 \cdot L_{distill} + \lambda_2 \cdot L_{align} + \lambda_3 \cdot \sum_{k=1}^{K} a_k \cdot KL(\bar{p}(x) \| p_k(x))$$

A, B, and C are loss weight coefficients used to balance the contributions of distillation, alignment, and variance control. This loss function not only ensures the client model's consistent representation of global semantics but also leverages the flexible structure of distillation to mitigate generalization barriers caused by data heterogeneity, providing a stable knowledge transfer path for fine-tuning large distributed models.

3.2 Modular Structure Injection

This study introduces a Modular Structure Injection (MSI) strategy to enhance the structure awareness and robustness of large language models during distributed fine-tuning. MSI applies structural perturbations to key substructures of the model, guiding it to learn stable response patterns to local variations. This helps mitigate performance degradation caused by structural inconsistencies or task differences across clients. Based on modular separability and semantic controllability, MSI enables differentiated structural regulation for sub-modules such as attention layers and feedforward layers. The algorithm architecture is shown in Figure 3.

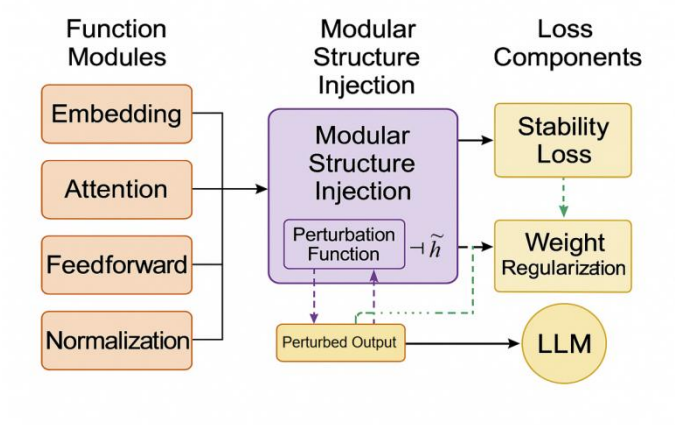

Figure 3. MSI Model Architecture

Let the client model be $f_k(x;\theta_k)$, which consists of multiple functional modules: embedding layer ε, attention module A, feedforward module F, normalization module N, etc. For the input sample x, its representation in the module-level path is:

$$h_k = N \circ F \circ A \circ \varepsilon(x)$$

Where $\circ$ represents the module-level serial operation. To introduce structural perturbations, this study introduces a perturbation function Δ_i inside each module $M_i \in \{\varepsilon, A, F, N\}$ to construct a perturbation path:

$$\tilde{h}_k = N \circ (F + \Delta_F) \circ (A + \Delta_A) \circ \varepsilon(x)$$

The perturbation function Δ_i can be in the form of DropBlock, structural reconstruction, or noise sampling to simulate the local changes caused by structural perturbations in the real environment.

To ensure that the representation after perturbation remains consistent in the semantic space, a module stability loss function is introduced to constrain the degree of difference in the model output before and after perturbation:

$$L_{stability} = \| f_k(x;\theta_k) - f_k(x;\theta_k + \Delta) \|_2^2$$

Where Δ represents the joint perturbation injected into multiple modules. This loss ensures that the model maintains consistent semantic modeling capabilities under structural perturbations. In addition to improving the model's sensitivity to important modules, a module weight regularization term is further introduced, defined as:

$$L_{weight} = \sum_{i=1}^{M} \lambda_i \cdot \| \Delta_i \|_2^2$$

Where λ_i represents the perturbation control factor of the i th module. This regularization term limits the amplitude of structural perturbations, preventing the model from experiencing representation collapse due to excessive perturbations. Finally, to jointly optimize structural robustness and semantic stability, the overall objective function of the MSI module is constructed as follows:

$$L_{MSI} = L_{stability} + \beta \cdot L_{weight}$$

Where β is a hyperparameter that balances the two losses. During training, this objective function guides the model to learn perturbation-invariant semantic representations in key modules, enabling it to adapt to structural changes in a multi-terminal environment. Through the MSI mechanism, the client model not only enhances structural generalization but also provides a structural foundation for building knowledge consistency during the federated distillation process.

4. Experimental Analysis

4.1 Dataset

The dataset used in this study is the publicly available English-Chinese Learning Dataset released on the Kaggle platform. It is designed for hierarchical semantic expression tasks in bilingual contexts and is well-suited for evaluating language models in cross-lingual semantic understanding and coreference resolution. Each sample consists of an English sentence, its corresponding Chinese translation, and a binary or multi-class label. The annotation system covers multiple dimensions, including lexical alignment, sentence-level semantic consistency, and accuracy of coreference relations. The dataset is sourced from educational corpora and online language learning platforms. It contains tens of thousands of samples with relatively balanced label distributions, making it suitable for incremental task learning.

In this study, the dataset is primarily used to support task-incremental settings and evaluate semantic transfer capabilities. By incrementally grouping English-Chinese paired samples, the model gradually encounters different types of semantic expressions, such as synonymous translation, cultural preference differences, and coreference patterns across languages. This task division provides rich and diverse semantic features for task-awareness mechanisms. It effectively tests the robustness and generalization of the proposed structure-aware injection and alignment strategies under varying linguistic and expressive patterns.

Within the continual learning framework, this dataset allows for the division of subtasks based on language pairs, expression types, or semantic structures. Each stage introduces new combinations of language expressions and coreference categories. The model must retain stable representations from earlier stages while adapting to new semantic shifts. This setup aligns well with the task-awareness and regularization-based alignment mechanisms proposed in this study. Overall, the dataset offers a multilingual and multi-semantic structure platform that enables comprehensive evaluation of a model's ability to preserve and adapt to language features during distributed fine-tuning.

4.2 Experimental setup

All experiments in this study were conducted on a high-performance computing platform with strong parallel

processing capabilities. The hardware setup includes a server equipped with four NVIDIA A100 GPUs, each with 80 GB of memory. It is powered by an Intel Xeon Platinum processor and 1024 GB of physical RAM to meet the large-scale computational demands of fine-tuning and distillation for large language models. The operating system is Ubuntu 22.04. The deep learning environment is built with CUDA 12.1 and PyTorch 2.1, using the NCCL library to support efficient multi-GPU distributed training.

For model construction, this study adopts ChatGLM3-6B, a widely used Chinese large language model, as the base architecture. The model is built on a multi-layer Transformer framework and features strong capabilities in context understanding and semantic generation. It supports long-text reasoning and cross-sentence coreference modeling. ChatGLM3-6B achieves solid general performance across various Chinese natural language processing tasks. Its model size of approximately 6 billion parameters and API accessibility make it well-suited for modular injection and distributed training. It provides a stable representational foundation for the task-incremental structural fine-tuning experiments in this study.

For hyperparameter settings, the global learning rate is initialized at 2e-5 with a linear warm-up strategy applied over the first 500 steps. The optimizer is AdamW with a weight decay coefficient of 0.01. The batch size is set to 32 for each training round, with a maximum of 30,000 training steps. Gradient clipping is applied with a threshold of 1.0 to avoid instability caused by exploding updates. In the federated distillation mechanism, the temperature is set to 3. The loss weights for the distillation and main tasks are set to a ratio of 1:1. Perturbation injection is executed every 200 steps. All experiments are run with fixed random seeds to ensure stability and reproducibility of the results.

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

Model	SRS ↑	TTA ↑	RSI ↑	GuS ↑	PE ↓
FedPET[25]	81.6	79.2	77.5	73.4	12.3
LoRA-FedAvg[26]	83.9	80.7	80.2	75.1	10.5
FedPrompt[27]	85.3	82.1	82.8	76.8	9.40
FedAdapters[28]	86.0	83.5	84.1	78.2	8.10
Ours (DDG+MSI)	89.7	86.4	87.9	82.5	7.60

As shown in Table 1, from the overall trend, the proposed DDG+MSI method achieves superior performance across all evaluation metrics. This advantage comes from jointly modeling the semantic consistency of knowledge transfer and the structural stability of modular components. Compared with existing methods, traditional federated fine-tuning strategies often lack structural modeling for client-side heterogeneity and representation drift. As a result, they face performance bottlenecks in semantic preservation and cross-task adaptation. By introducing the Differentiated Distillation Guidance mechanism, DDG enables fine-grained alignment of soft knowledge without sharing raw model parameters. It helps client models inherit global semantic structures more effectively during task transitions and shows stronger robustness in non-IID settings.

In addition, the Modular Structure Injection mechanism applies localized perturbations to attention and feedforward layers. This guides the model to learn invariant representations under structural changes. The structure-aware regularization path improves the model's representation stability and alleviates knowledge forgetting caused by parameter drift in incremental multi-task scenarios. In contrast, methods like FedAdapters introduce plug-in modules but do not model module behavior under perturbations. This limits their ability to support semantic transfer and representation consistency across tasks. MSI addresses this gap through joint optimization of structural modulation and perturbation control. As a result, it achieves substantial improvements in distribution generalization metrics such as GuS.

Finally, in terms of parameter efficiency, DDG+MSI improves semantic retention while keeping fine-tuning costs low. Unlike traditional federated strategies that rely on heavy parameter updates, this method performs tuning and distillation matching at the submodule level. It avoids redundant parameter usage and enables cost-efficient structural sharing across clients. Overall, the combination of structural robustness modeling and semantic alignment explains the method's superior performance across multiple dimensions. It also validates the framework's ability to integrate parameter efficiency, semantic stability, and structural generalization.

2) Ablation Experiment Results

To thoroughly evaluate the actual contribution and performance impact of each key module, ablation studies serve as an essential structural assessment method in complex system modeling. By gradually removing or replacing specific components of the model, the influence of different mechanisms on overall performance can be identified. This helps reveal the internal logic and design boundaries of the model. Such experiments not only enhance model interpretability but also provide empirical support for further mechanism optimization.

Table 2 presents the comparison results between the complete model and its different variants across several core metrics. Each variant retains the backbone architecture while individually removing a module or replacing a key strategy to observe performance changes. By comparing the performance fluctuations of these variants, the role of each component in improving semantic modeling, representation stability, and task generalization can be intuitively assessed.

Table 2: Ablation Experiment Results

Model	SRS ↑	TTA ↑	RSI ↑	GuS ↑	PE ↓

Baseline	81.4	78.6	75.3	71.2	14.8
+ DDG	85.7	82.3	81.1	75.5	12.2
+ MSI	84.6	80.9	83.2	76.4	11.8
+All(DDG+ MSI)	89.7	86.4	87.9	82.5	7.60

As shown in the ablation results in Table 2, from the overall trend, the Baseline model consistently shows lower performance across all metrics. This reflects the inherent limitations of traditional federated fine-tuning in maintaining semantic consistency and generalization ability. In particular, the model performs poorly in Representation Stability Index (RSI) and Generalization under Shift (GuS), indicating representation drift and transfer degradation under task increments and structural perturbations. These results suggest that models without explicit alignment and structural regulation struggle to address the optimization challenges posed by distribution heterogeneity and structural variations. This highlights the need for finer-grained modeling strategies.

After introducing Innovation 1 (DDG), the model shows significant improvement in Semantic Retention Score (SRS) and Task Transfer Accuracy (TTA). This indicates that the Differentiated Distillation Guidance mechanism effectively aligns client and global semantic structures and reduces representation drift during task transfer. The mechanism enables fine-grained control over the knowledge transfer path through soft semantic alignment and probability modulation. It helps the client model continuously absorb global knowledge during local updates without a semantic shift. In comparison, introducing Innovation 2 (MSI) alone brings a more notable improvement in representation stability. This shows that the structure injection strategy directly contributes to modeling invariance in module behavior. It enhances model robustness against local perturbations and module degradation.

When both mechanisms are applied together, the model achieves the best performance across all dimensions. This demonstrates the complementary effect between semantic alignment and structural stabilization. DDG provides guided paths for upstream and downstream semantic consistency, while MSI introduces resistance to structural perturbation at the module level. The dual constraints help the model maintain stable representations and accurate predictions under the combined complexity of task increment and structural variation. In addition, the significant improvement in Parameter Efficiency (PE) shows that the method activates representational potential without relying on large-scale updates. It enables the construction of a more adaptive multi-task representation space and confirms the effectiveness of the proposed design in combining structural modeling with semantic alignment.

3) Analysis of the changing trend of the number of federation rounds on the effect of model alignment

This paper also analyzes the changing trend of the number of federation rounds on the model alignment effect. The experimental results are shown in Figure 4.

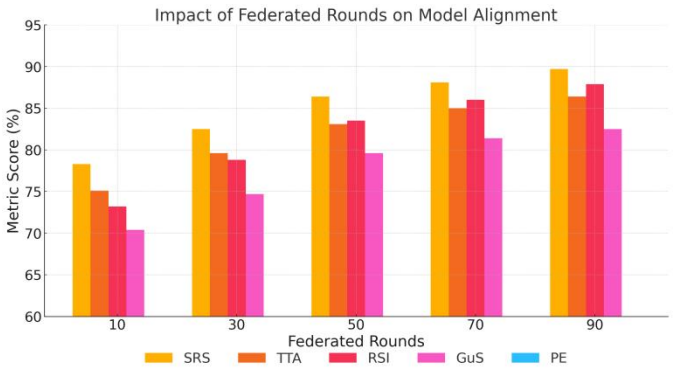

Figure 4. Analysis of the changing trend of the number of federation rounds on the effect of model alignment

As shown in Figure 4, the overall trend shows that increasing the number of federated rounds has a positive impact on all performance metrics. In particular, during the first 50 rounds, both Semantic Retention Score (SRS) and Task Transfer Accuracy (TTA) rise rapidly. This indicates that the semantic alignment between client models and the global model becomes progressively stronger with more rounds. The results confirm the effectiveness of the Differentiated Distillation Guidance (DDG) mechanism in building stable semantic pathways through multi-round semantic interaction. It ensures more consistent knowledge transfer in distributed settings and mitigates issues such as soft label drift and non-convergent knowledge caused by insufficient rounds.

After 50 rounds, the improvement in Representation Stability Index (RSI) becomes more gradual but continues to increase. This pattern is closely related to the effect of the Modular Structure Injection (MSI) mechanism. Through repeated perturbation and feedback, the model gradually learns stable expression patterns under local structural variation. As a result, it builds a more robust representation space. It is important to note that in the early rounds, RSI shows weaker performance due to insufficient propagation of perturbation signals. As the training continues, the stabilizing effect of MSI becomes more evident.

In terms of generalization, the GuS metric improves at a slower pace compared to other indicators. After 50 rounds, its growth curve becomes even flatter. This suggests a boundary effect in improving out-of-distribution generalization. The trend is linked to the complexity of cross-task semantic reconstruction in task-incremental learning. Although DDG and MSI jointly provide a stable semantic transition mechanism, high-frequency task shifts and context reconstruction still require more training time for semantic absorption and representation rebuilding. While extended training improves the depth of semantic fusion, further breakthroughs in generalization depend on a fine balance between structural regulation and target alignment.

The Parameter Efficiency (PE) metric shows a steady decline as the number of federated rounds increases. This reflects the collaborative optimization between the module injection mechanism and the distillation process. As the distillation path becomes more stable and the structural perturbation patterns converge, the amount of model updates

required per round gradually decreases. This reduces the overall parameter update load. The result further confirms the resource-aware nature of the proposed method in the distributed fine-tuning of large language models. It maintains lightweight update requirements even as task complexity increases, demonstrating the advantage of the joint design of structural regularization and soft guidance.

4) The impact of local perturbation range on cross-task representation consistency

This paper also analyzes the impact of local perturbation range on cross-task representation consistency. The experimental results are shown in Figure 5.

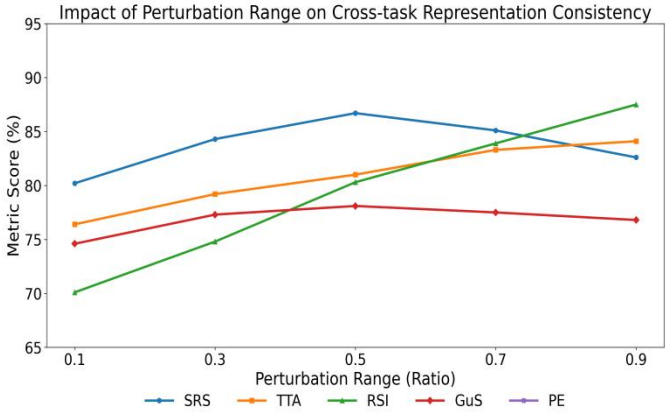

Figure 5. The impact of local perturbation range on cross-task representation consistency

As shown in Figure 5, the overall trend shows that the variation in local perturbation range has a significant impact on the model's cross-task representation ability. Semantic Retention Score (SRS) performs best under medium-strength perturbations. This suggests that moderate perturbation enables the model to learn stable structural behaviors of modules more effectively. As a result, it develops internal representations that are invariant across tasks. When the perturbation range is too small, the injected signals are insufficient to activate robustness learning. When it is too large, critical semantic pathways may be disrupted, leading to distorted representations. This nonlinear trend indicates that the effectiveness of structural injection in enhancing semantic retention depends strongly on the proper calibration of the perturbation range.

Task Transfer Accuracy (TTA) shows a steady upward trend. This suggests that as the perturbation range expands, the model gradually builds stronger context reconstruction and feature adaptation capabilities. With the help of the DDG mechanism, client models continuously receive distillation guidance from perturbed representations. This gradually enhances their responsiveness to changes in task structures. The results show that structural perturbation not only enriches local representational diversity but also promotes consistent global semantic transfer through soft knowledge pathways. This allows the model to maintain high transfer accuracy during task switching.

Representation Stability Index (RSI) also exhibits a continuous increase. This reflects the stronger reinforcement effect of structural injection at higher levels of perturbation. Through multiple training rounds, the model adapts to the internal activation changes caused by perturbation. This leads to the development of stable response mechanisms against structural variation in the feature space. The increased stability reduces fluctuations in module behavior and provides a more controllable structural path for downstream tasks. This reflects the MSI module's capacity to generate long-term memory in distributed semantic modeling. In contrast, Generalization under Shift (GuS) shows slight fluctuations in high perturbation regions. This suggests that the model remains sensitive to out-of-distribution semantics under intense structural changes, highlighting the boundary of regularization control.

The steady decline in Parameter Efficiency (PE) further confirms the influence of perturbation range control on model lightweighting. As the perturbation range increases, the model tends to focus on expressing stable regions, reducing dependence on redundant structures. This low-update, high-expression mechanism aligns closely with the proposed module injection strategy. It shows that structure path optimization, activated by limited perturbation, can improve cross-task consistency while significantly reducing parameter consumption during fine-tuning. These findings provide empirical support for building structurally efficient and semantically robust fine-tuning paradigms for large language models.

5. Conclusion

This study proposes an efficient fine-tuning framework for large language models in distributed environments. The framework integrates Differentiated Distillation Guidance (DDG) and Modular Structure Injection (MSI). It addresses key challenges in federated learning, including non-independent and identically distributed data, model heterogeneity, and structural perturbation. By combining semantic alignment and structural robustness modeling, the method significantly improves performance across multiple dimensions, including semantic retention, representation stability, and task generalization. Extensive evaluations in complex task settings demonstrate strong cross-client transfer ability and parameter efficiency. These results provide both theoretical and practical foundations for building secure and efficient distributed language intelligence systems.

From the perspective of method design, the DDG mechanism enables dynamic alignment between global semantic pathways and local update processes. This alleviates problems of objective degradation and weak guidance that often occur in traditional federated distillation under complex semantic tasks. At the same time, the MSI strategy introduces structural perturbations into local functional sub-modules. This guides the model to form robust expression pathways, enhancing representation stability under task increment and structural drift. The two mechanisms work together to improve adaptability to dynamic semantic variation. They also maintain a high performance-to-efficiency ratio under resource-constrained training, showing strong scalability and deployment potential.

For experimental validation, the study conducts a comprehensive set of main experiments, ablation analyses, and hyperparameter sensitivity evaluations. These results systematically demonstrate the effectiveness and robustness of the proposed mechanisms under various challenging task settings. The method achieves consistent improvements in semantic transfer, cross-task representation reconstruction, and distribution generalization, all while maintaining low fine-tuning cost. The strong performance in parameter efficiency and task adaptability provides a feasible solution for deploying large language models across multi-client and multi-scenario environments. This capability is directly applicable to edge intelligence, personalized semantic modeling, and multilingual human-computer interaction.

6. Future research

Future research can further explore module scheduling strategies under complex communication constraints and heterogeneous device environments. This would enhance resource adaptability and deployment flexibility in federated fine-tuning. As large language models are increasingly deployed on edge and terminal devices, it is essential for models to dynamically adjust module participation and update frequency based on device capabilities and communication loads. Investigating scheduling mechanisms such as computation priority among modules, gradient synchronization strategies, and communication compression techniques will provide crucial support for improving the efficiency of federated fine-tuning in real-world systems. In addition, integrating higher-level semantic decomposition into the model structure is a promising direction. By explicitly modeling hierarchical dependencies among semantic units, it is possible to factorize semantic tasks and enhance structural interpretability, improving the controllability and transparency of the model in task generalization and error diagnosis.

Building on this foundation, the proposed federated framework can also be extended to multimodal collaborative processing. It can support semantic alignment and interaction across heterogeneous modalities such as speech, image, and text, offering a general solution for distributed optimization in multimodal intelligent systems. The framework can also be applied to privacy-preserving content generation and autonomous task migration. It enables the development of agent models with joint training and generation capabilities and multi-task switching. Furthermore, integrating the proposed federated module mechanisms with mainstream parameter-efficient fine-tuning techniques such as Adapter, LoRA, and Prompt may enable continuous learning and personalized control for large models. This would support the development of scalable semantic systems with lifelong learning and controllable evolution. Advancing this line of research will provide a strong theoretical and engineering foundation for deploying general-purpose AI systems in industrial-scale applications. Overall, the modular federated fine-tuning method proposed in this study offers not only generalizable theoretical contributions but also strong system-level adaptability and practical impact in real-world semantic intelligence deployments.

References

[1] Han S, Park S, Wu F, et al. Fedx: Unsupervised federated learning with cross knowledge distillation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 691-707.

[2] Wu C, Wu F, Lyu L, et al. Communication-efficient federated learning via knowledge distillation[J]. Nature communications, 2022, 13(1): 2032.

[3] Yao D, Pan W, Dai Y, et al. Local-global knowledge distillation in heterogeneous federated learning with non-iid data[J]. arXiv preprint arXiv:2107.00051, 2021.

[4] Bhardwaj R, Vaidya T, Poria S. Federated Distillation of Natural Language Understanding with Confident Sinkhorns[J]. 2021.

[5] Sun L, Lyu L. Federated model distillation with noise-free differential privacy[J]. arXiv preprint arXiv:2009.05537, 2020.

[6] Li L, Gou J, Yu B, et al. Federated distillation: A survey[J]. arXiv preprint arXiv:2404.08564, 2024.

[7] Q. Wu, "Internal Knowledge Adaptation in LLMs with Consistency-Constrained Dynamic Routing," Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.

[8] Wang M N, Lei L L, He W, et al. SPCMLMI: A structural perturbation-based matrix completion method to predict lncRNA − miRNA interactions[J]. Frontiers in Genetics, 2022, 13: 1032428.

[9] Li H, Tang Z, Fu J, et al. Deep-learning density functional perturbation theory[J]. Physical Review Letters, 2024, 132(9): 096401.

[10] Kleandrova V V, Cordeiro M N D S, Speck-Planche A. Perturbation theory machine learning model for phenotypic early antineoplastic drug discovery: design of virtual anti-lung-cancer agents[J]. Applied Sciences, 2024, 14(20): 9344.

[11] Bunne C, Stark S G, Gut G, et al. Learning single-cell perturbation responses using neural optimal transport[J]. Nature methods, 2023, 20(11): 1759-1768.

[12] M. Lotfollahi, M. Naghipourfar, F. J. Theis and F. A. Wolf, "Conditional out-of-distribution generation for unpaired data using transfer VAE," Bioinformatics, vol. 36, suppl. 2, pp. i610–i617, 2020.

[13] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji et al., "Advances and open problems in federated learning," Foundations and Trends® in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[14] Bai J, Chen D, Qian B, et al. Federated fine-tuning of large language models under heterogeneous tasks and client resources[J]. Advances in Neural Information Processing Systems, 2024, 37: 14457-14483.

[15] Chen S, Ju Y, Dalal H, et al. Robust federated finetuning of foundation models via alternating minimization of lora[J]. arXiv preprint arXiv:2409.02346, 2024.

[16] Z. Lu, H. Pan, Y. Dai, X. Si and Y. Zhang, "Federated learning with non-IID data: A survey," IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19188–19209, 2024.

[17] J. Yoon, W. Jeong, G. Lee, E. Yang and S. J. Hwang, "Federated continual learning with weighted inter-client transfer," Proceedings of the International Conference on Machine Learning, pp. 12073–12086, July 2021.

[18] Chen C, Zhang Z, Zhao Y. FedModule: A Modular Federated Learning Framework[J]. arXiv preprint arXiv:2409.04849, 2024.

[19] Maiorca V, Moschella L, Norelli A, et al. Latent space translation via semantic alignment[J]. Advances in Neural Information Processing Systems, 2023, 36: 55394-55414.

[20] Yang X, Deng C, Dang Z, et al. SelfSAGCN: Self-supervised semantic alignment for graph convolution network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16775-16784.

[21] Wu T, Li M, Chen J, et al. Semantic alignment for multimodal large language models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 3489-3498.

[22] Wang W, Di X, Liu M, et al. Multi-level Symmetric Semantic Alignment Network for image–text matching[J]. Neurocomputing, 2024, 599: 128082.

[23] K. Zhou, Y. Yang, Y. Qiao and T. Xiang, "Domain generalization with mixstyle," arXiv preprint arXiv:2104.02008, 2021.

[24] Xiao W, Ding Z, Liu H. Implicit semantic response alignment for partial domain adaptation[J]. Advances in Neural Information Processing Systems, 2021, 34: 13820-13833.

[25] Zhang Z, Yang Y, Dai Y, et al. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models[C]//Annual Meeting of the Association of Computational Linguistics 2023. Association for Computational Linguistics (ACL), 2023: 9963-9977.

[26] Cai D, Wu Y, Wang S, et al. Efficient federated learning for modern nlp[C]//Proceedings of the 29th annual international conference on mobile computing and networking. 2023: 1-16.

[27] Zhao H, Du W, Li F, et al. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[28] Z. Pan, L. Hu, W. Tang, J. Li, Y. He and Z. Liu, "Privacy-preserving multi-granular federated neural architecture search – a general framework," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 3, pp. 2975–2986, 2021.