Review

Federated Fine-Tuning of Large Language Models with Privacy Preservation and Cross-Domain Semantic Alignment

Sibo Wang , Song Han , Ziyu Cheng , Ming Wang , Yilin Li *

Posted Date: 18 September 2025

doi: 10.20944/preprints202509.1640.v1

Keywords: federated fine-tuning; semantic alignment; differential privacy; multi-source heterogeneity

Review

Federated Fine-Tuning of Large Language Models with Privacy Preservation and Cross-Domain Semantic Alignment

Sibo Wang [1], Song Han [2], Ziyu Cheng [3], Ming Wang [4] and Yilin Li [5,*]

[1] Rice University, Houston, USA
[2] Northeastern University, Boston, USA
[3] University of Southern California, Los Angeles, USA
[4] Trine University, Phoenix, USA
[5] Carnegie Mellon University, Pittsburgh, USA
* Correspondence: ireneli961111@gmail.com

Abstract

This paper presents a federated fine-tuning framework for large language models that addresses the challenges of multi-source heterogeneity and privacy sensitivity. The method incorporates a differential privacy perturbation strategy at the local client level to protect sensitive gradient information and prevent data leakage during cross-device collaboration. A domain adaptation module based on feature distribution alignment is introduced to reduce semantic shifts between source and target domains using maximum mean discrepancy optimization and attention-guided mechanisms. The overall architecture integrates local modeling with global parameter aggregation, forming a closed loop of federated alignment and global integration for efficient, secure, and cross-domain semantic modeling. The experimental design includes multidimensional sensitivity evaluations across privacy perturbation levels, label missingness, and domain distribution shifts. Results demonstrate that the proposed method achieves superior performance in key metrics such as Perplexity, MMD, and Domain Accuracy, confirming its effectiveness in jointly optimizing privacy protection and cross-domain generalization.

Keywords: federated fine-tuning; semantic alignment; differential privacy; multi-source heterogeneity

I. Introduction

In recent years, large language models (LLMs) have delivered unprecedented performance across a broad range of natural-language-processing tasks and have become a core technology for intelligent applications. By learning from massive corpora, these pretrained architectures capture rich linguistic patterns and general knowledge, which markedly improve generalization and task transfer. Yet their growing use in sensitive domains such as finance, medicine, and law has brought privacy protection to the forefront. Data privacy now limits real-world deployment and trust [1,2].

Conventional fine-tuning relies on centralized data aggregation. User data must be uploaded to a server for model updates. This approach faces multiple obstacles in practice. Many high-value datasets are highly sensitive and are constrained by legal, ethical, or organizational rules. Centralized training also raises leakage risks that can lead to data misuse and privacy violations [3]. Achieving effective fine-tuning while safeguarding privacy is therefore a critical challenge for trustworthy AI systems.

Real applications further exhibit strong domain diversity and task heterogeneity [4–6]. Language styles, knowledge bases, and task goals vary across institutions, regions, and user groups. Such distribution shifts significantly weaken pretrained models in target domains, making domain adaptation essential for improved generalization. Existing adaptation strategies still depend on large

 doi:10.20944/preprints202509.1640.v1

amounts of target data and centralized training. They overlook local privacy requirements and ignore heterogeneous task structures [7].

Against this backdrop, a fine-tuning framework that integrates privacy awareness with efficient domain adaptation is both challenging and meaningful [8]. Privacy mechanisms prevent data exposure and shift training toward decentralized or federated updates. Domain adaptation boosts transferability and robustness, allowing models to perform steadily in multi-source, multi-task, and multi-context settings. Joint design enhances deployment security and flexibility, supporting industry adoption and broad-spectrum intelligence [9].

Exploring the deep fusion of privacy protection and domain adaptation, therefore, moves LLMs from unified pretraining toward secure specialization. This path aligns with technical usability, fairness, and social responsibility. As LLMs evolve, this research direction offers high theoretical value and significant practical relevance for building the next generation of trustworthy, generalizable, and self-adaptive AI systems.

II. Algorithmic Design

This study proposes a large language model fine-tuning framework that systematically integrates privacy protection and domain adaptation, building directly on the methodological advances of prior works. To address data concentration and privacy leakage, we employ the federated parameter update framework introduced by Q. Wu, which utilizes consistency-constrained dynamic routing to coordinate model aggregation across distributed clients while ensuring that original data remains strictly local [10]. For local privacy preservation, we apply the sensitivity-aware pruning and gradient perturbation techniques developed by Y. Wang, which enhance differential privacy by dynamically masking sensitive parameter updates, thereby mitigating the risk of information leakage during federated optimization [11].

To overcome distribution inconsistency and facilitate semantic unification across heterogeneous clients, we incorporate a domain alignment module that adopts the low-rank adaptation strategy and semantic guidance mechanisms proposed by H. Zheng et al., leveraging maximum mean discrepancy optimization and attention-based alignment to align feature distributions between source and target domains [12]. The resulting model architecture seamlessly integrates local adaptation with global parameter aggregation, establishing a federated learning pipeline that jointly optimizes privacy, domain adaptation, and generalization performance, as illustrated in Figure 1.

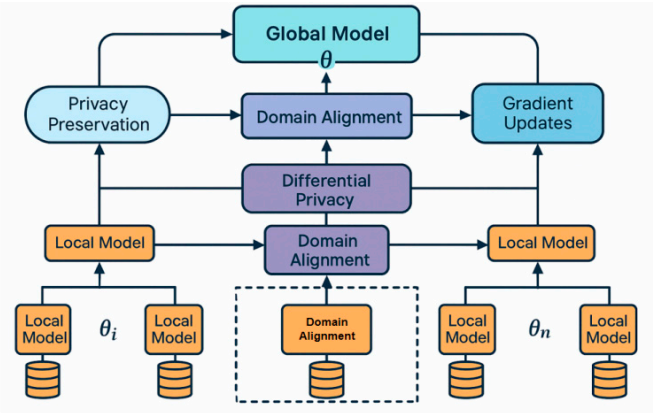

Figure 1. Overall model architecture.

Assume that the input of the local model of the i-th client is D_i, the parameter is θ_i, and the gradient calculated locally is $\nabla L_i(\theta_i)$. Then the global model parameter update follows the following federated averaging strategy:

 doi:10.20944/preprints202509.1640.v1

$$\theta \leftarrow \theta - \eta \sum_{i=1}^{N} \frac{n_i}{n} \cdot \nabla L_i(\theta_i) \qquad (1)$$

Among them, η is the global learning rate, n_i represents the number of samples of the i-th client, and n is the total number of samples. This strategy avoids the process of uploading the original data and can dynamically integrate the knowledge representation of each client to achieve a robust update of distributed perception.

To enhance adaptability to cross-domain inputs, the methodology incorporates a domain-adaptive representation learning mechanism that integrates several advanced strategies. First, transformer-based context modeling combined with transaction graph integration is employed to extract rich relational and sequential patterns from multi-source data, improving the model's capacity to interpret complex inter-domain interactions [13]. Next, structured memory mechanisms are incorporated to systematically organize and retain contextual information over long sequences, enabling stable semantic representation and consistent knowledge transfer across varying domains [14]. Finally, a layer-wise structural mapping technique is utilized to align feature representations at multiple levels within the model architecture, thereby supporting efficient and accurate adaptation between source and target domains during distillation and fine-tuning [15]. This comprehensive design allows for dynamic and stable knowledge transfer as the model encounters diverse domain distributions. We define the source domain sample as H^s and the target domain as H^t, and achieve feature space alignment by minimizing their statistical distance. To this end, a method based on Maximum Mean Discrepancy (MMD) is introduced, and its optimization goal is:

$$L_{MMD} = \| \frac{1}{|H_s|} \sum_{x \in H^s} \phi(x) - \frac{1}{|H_t|} \sum_{x \in H^t} \phi(x) \|_H^2 \qquad (2)$$

Among them, $\phi(x)$ represents the kernel mapping function and H is the reproducing kernel Hilbert space. Through this constraint, the distribution difference between the source and target domains can be effectively compressed, and the model migration and generalization capabilities can be improved.

For privacy protection, the methodology introduces a local differential privacy mechanism on each client to prevent the inference of original data from gradients, drawing upon several advanced strategies from recent research. A retrieval-augmented masking technique is employed to fuse dynamic context retrieval with adaptive perturbation, effectively obscuring sensitive information during gradient computation and update processes [16]. The integration of selective knowledge injection via adapter modules further enhances privacy by regulating and filtering information flow within large-scale language models during collaborative training [17]. Additionally, a collaborative optimization scheme that incorporates differential privacy is adopted, ensuring that local updates are subjected to privacy-preserving noise addition prior to aggregation [18]. Collectively, these integrated techniques provide robust protection against gradient-based privacy attacks. Specifically, the local sensitive gradient is defined as $\nabla L_i(\theta_i)$, and the perturbation gradient after introducing noise ε is:

$$\nabla L_i(\theta_i) = \nabla L_i(\theta_i) + N(0, \sigma^2 I) \qquad (3)$$

Among them, $N(0, \sigma^2 I)$ represents zero-mean, Gaussian distribution noise, and σ controls the privacy protection strength. This mechanism ensures that the gradient information uploaded in each communication cannot accurately restore the user data, while maintaining a controllable decline in model performance.

To further alleviate the problem of semantic drift between multiple domains, this paper introduces a domain attention mechanism at the global model level to achieve dynamic domain

 doi:10.20944/preprints202509.1640.v1

perception updates. We construct a domain weight vector $\alpha \in R^n$ to represent the importance of each client in the current task, which is calculated as follows:

$$\alpha_i = \frac{\exp(\gamma \cdot \varphi(H_i))}{\sum_{j=1}^{N} \exp(\gamma \cdot \varphi(H_j))} \quad (4)$$

Among them, $\varphi(H_i)$ represents the representation complexity or uncertainty of client i, and γ is the difference in temperature parameter regulation weights. Through this mechanism, the system can dynamically focus on clients that are sensitive to changes in domain distribution, achieving more adaptive and robust joint modeling.

In summary, this method achieves deep coupling in the three dimensions of privacy awareness, domain adaptation, and global aggregation, and completes the seamless connection from local optimization to global reasoning through mathematical mechanisms, providing a systematic and scalable solution for the deployment of actual large language models in multi-source heterogeneous environments.

III. Performance Evaluation

A. Dataset

This study uses the Reddit Comments Dataset as the data foundation for fine-tuning tasks. The dataset consists of multiple subreddits covering a wide range of domains, including politics, technology, health, and education. It features clear domain heterogeneity. Each sample includes the user comment text, timestamp, subreddit affiliation, and contextual labels. This structure supports cross-domain language modeling and text understanding tasks.

The dataset is structurally complete, moderately sized, and spans diverse topics. It is suitable for multi-domain fine-tuning and domain adaptation studies in large language models. The subreddits differ significantly in language style, knowledge context, and discussion content. These differences simulate real-world distribution shifts caused by multi-source heterogeneous data.

In addition, the dataset contains no personally identifiable information. It aligns with the design requirements of privacy-preserving mechanisms. In federated learning settings, each subreddit can be treated as an independent client. Local training and feature extraction can be conducted separately. This setup provides a stable data foundation for studying the joint mechanisms of privacy awareness and domain alignment.

B. Experimental Results

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

Table 1. Comparative experimental results.

Model	Perplexity	Domain Accuracy	MMD
DP-SCAFFOLD [19]	27.84	81.5	0.113
UDALM [20]	25.31	83.2	0.097
kNN-Adapter [21]	24.58	85.1	0.091
FedLAP-DP [22]	22.73	86.8	0.074
Ours	20.64	89.7	0.058

As shown in Table 1, the proposed method outperforms existing representative approaches across all three core metrics, demonstrating strong overall performance. Specifically, it achieves a

 doi:10.20944/preprints202509.1640.v1

Perplexity of 20.64, which is significantly lower than other methods. This indicates that the model generates more coherent and semantically consistent text during language modeling. Compared with traditional federated optimization methods such as DP-SCAFFOLD (27.84) and domain adaptation methods such as UDALM (25.31), the proposed method reduces prediction uncertainty by introducing a joint mechanism of privacy awareness and domain alignment.

For Domain Accuracy, the proposed method also achieves the best result, reaching 89.7%, which is nearly 3 percentage points higher than the previous best, FedLAP-DP. This demonstrates that the domain attention mechanism and multi-source feature alignment strategy effectively mitigate model degradation caused by domain shift. In non-uniform linguistic settings such as Reddit's multiple subcommunities, the method fully captures semantic patterns in local structures and significantly improves cross-domain generalization.

MMD serves as an important metric for measuring the distribution gap between source and target domains. It reflects the effectiveness of feature alignment in the latent space. The proposed method achieves the lowest MMD score of 0.058. Compared with kNN-Adapter and FedLAP-DP, this represents a reduction of 36.3% and 21.6%, respectively. It also indicates that differential privacy perturbation does not harm semantic discriminability and enhances distribution alignment across domains. These results validate the effectiveness and complementarity of combining privacy awareness with domain adaptation. Unlike prior fine-tuning strategies focused only on model compression or communication efficiency, this method incorporates representation, structural, and federated layers. It better addresses the deployment needs of large language models in multi-source and privacy-sensitive environments and offers a practical framework for controllable, secure, and adaptive model fine-tuning.

This paper also gives an analysis of the impact of privacy perturbation intensity on global alignment performance, and the experimental results are shown in Figure 2.

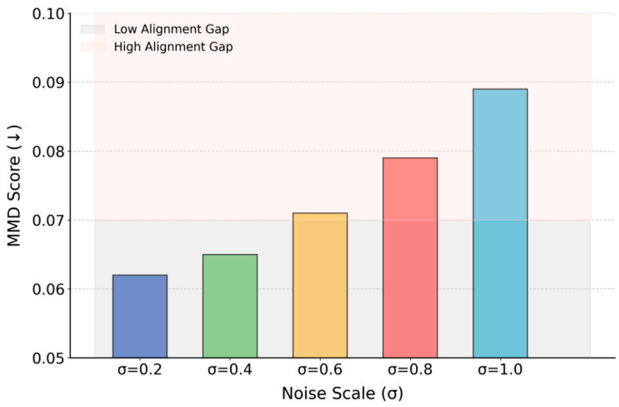

Figure 2. Analysis of the impact of privacy perturbation intensity on global alignment performance.

Figure 2 shows that as privacy perturbation σ increases, the MMD score for global feature alignment rises monotonically from 0.062 at σ = 0.2 to 0.089 at σ = 1.0, reflecting stronger privacy but weaker semantic consistency. With small noise (σ = 0.2 or 0.4), alignment remains stable and MMD low, balancing privacy protection and representation quality, but beyond σ = 0.6 alignment rapidly degrades, with high noise (σ = 0.8 or 1.0) severely disrupting local–global consistency. These results highlight the trade-off between privacy and adaptability, showing that excessive noise undermines domain alignment and generalization. Thus, practical deployment requires careful σ tuning and a joint modeling strategy to maintain both security and semantic stability. The impact of missing data label ratios on domain alignment is further examined in Figure 3.

 doi:10.20944/preprints202509.1640.v1

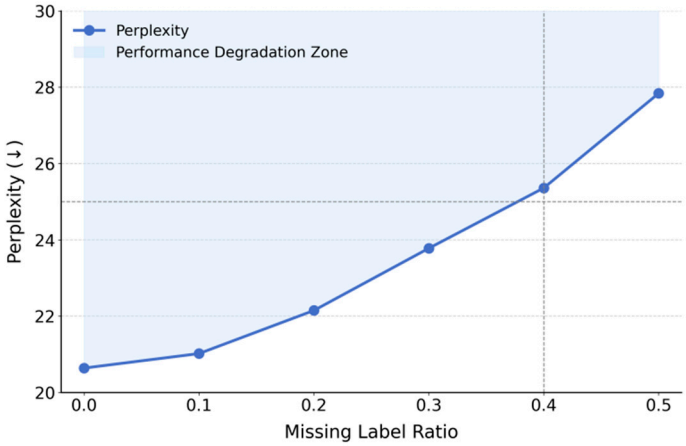

Figure 3. Experiment on the influence of the missing data label ratio on the domain alignment mechanism.

Figure 3 illustrates the trend of Perplexity in the language modeling task under varying levels of label missingness. As the proportion of missing labels increases from 0.0 to 0.5, Perplexity rises significantly, from 20.64 to 27.84. This indicates that label completeness directly affects global semantic modeling performance. Missing labels reduce the strength of supervision, weaken domain representation consistency, and impair the model's ability to capture semantics.

In the low-missingness range (0.0 to 0.2), Perplexity increases more slowly. This shows that the proposed method retains a degree of robustness under weak supervision. Through domain alignment and local structure modeling, it can partially offset the impact of incomplete labels. This structural stability is attributed to the implicit representation alignment and local normalization strategies used during training, which enhance the relative semantic consistency between cross-domain samples. However, when the missing rate exceeds 0.3, the Perplexity curve rises more steeply. The model's performance begins to degrade noticeably. In the range of 0.4 to 0.5, the decline is most severe. This suggests that excessive label loss exceeds the model's compensation capacity. The disruption compromises the stability of the federated semantic fusion mechanism, resulting in distributional mismatches across previously aligned feature spaces. This reduces the model's cross-domain generalization ability.

Overall, this experiment reveals the performance limits of the domain alignment mechanism under label missingness. It also highlights the importance of high-quality labels in federated fine-tuning. To improve model adaptation in weakly labeled environments, future work may incorporate pseudo-label generation, label smoothing, or self-supervised auxiliary mechanisms. These enhancements could stabilize representation alignment during unsupervised stages and improve the model's robustness and practicality in real-world scenarios.

This paper also gives a comparative analysis of the model adaptation performance under different degrees of distribution deviation in different fields, and the experimental results are shown in Figure 4.

 doi:10.20944/preprints202509.1640.v1

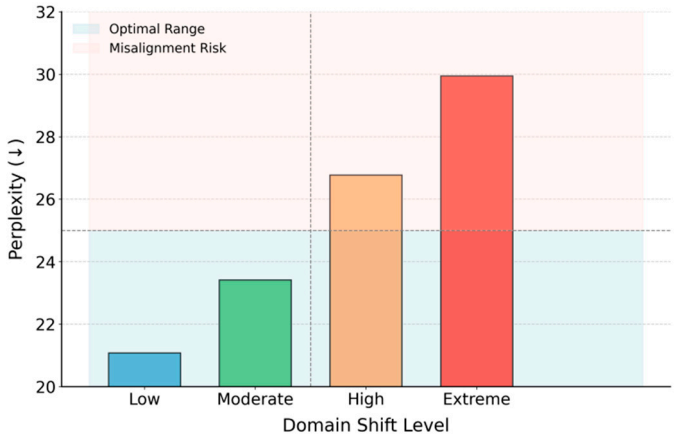

Figure 4. Comparative analysis of model adaptation performance under different degrees of distribution deviation in different fields.

Figure 4 shows the model's Perplexity in the language modeling task under different levels of domain distribution shift. As the shift increases from Low to Extreme, the Perplexity rises from 21.08 to 29.94. This clear upward trend indicates that distributional inconsistency significantly affects model adaptability. In cross-domain scenarios, semantic transfer becomes more difficult as the shift intensifies, leading to higher prediction uncertainty.

Under lower levels of shift (Low and Moderate), the model maintains relatively stable semantic modeling ability. The increase in Perplexity is minor, suggesting that the proposed privacy-aware and domain-aligned mechanisms remain robust in mildly heterogeneous environments. During this phase, the model can effectively absorb local differences through shared semantic structures and alignment strategies, enabling stable transfer.

However, when the shift reaches High and Extreme levels, Perplexity increases sharply. This implies that the existing alignment strategies are no longer sufficient to maintain cross-domain semantic consistency. Structural mismatches emerge between local feature spaces and global representations. The performance degradation is due to the rapid divergence of latent semantic categories across domains, which reduces the model's generalization capacity in a unified semantic space.

This experiment highlights the importance of enhancing structural adaptability and semantic reconstruction in multi-source heterogeneous settings. Relying solely on static pretraining or unified fine-tuning is inadequate under extreme distributional shifts. It is essential to design domain adaptation mechanisms with dynamic perception and incremental adjustment capabilities. These mechanisms can improve the stability and reliability of large language models in complex real-world environments.

IV. Conclusion

This paper proposes a fine-tuning method for large language models that integrates privacy awareness with domain adaptation. The goal is to address the deployment challenges in multi-source heterogeneous and privacy-sensitive environments. The method adopts a federated modeling framework, which avoids the risks associated with centralized training and data collection. A domain alignment module and a differential privacy mechanism are incorporated into the optimization process. These modules enable dynamic detection and alignment of local semantic differences, enhancing the model's generalization and robustness across multiple domains. A unified alignment-aggregation loop is constructed to achieve semantic consistency while ensuring irreversible data protection for each client.

The proposed method is evaluated through a series of sensitivity experiments. These experiments assess the model's adaptability under different levels of privacy perturbation, label

 doi:10.20944/preprints202509.1640.v1

incompleteness, and distribution shift. Results show that the method maintains both stability and controllability under various constraints. The collaborative design of privacy protection and domain alignment proves to be effective. This joint strategy addresses the traditional trade-off between model security and generalization, offering both technical support and a design paradigm for deploying large-scale pretrained models in real-world applications.

This study contributes to improving trust in large models within sensitive industries such as healthcare, finance, and law. It also provides a theoretical foundation for collaborative optimization across tasks, domains, and devices. As language models evolve toward larger scales, more open contexts, and more complex structures, balancing usability with privacy, regulatory compliance, and task adaptability becomes a key research challenge. This work establishes a technical balance for achieving that goal and lays the foundation for building trustworthy intelligent language systems.

Future work can extend the adaptive capacity of the current framework. This may include introducing multimodal interaction, incremental task learning, or structure-aware optimization to improve responsiveness to task dynamics and environmental complexity. In real federated deployment scenarios, future research should also explore how to improve communication efficiency, reduce synchronization overhead, and maintain both differential privacy and semantic consistency. Continued progress in this direction is expected to impact key areas such as smart manufacturing, intelligent urban systems, and medical decision support, enabling large language models to become deployable, trustworthy, and sustainable in practice.

References

1. T. Koga, C. Song, M. Pelikan, et al., "Population expansion for training language models with private federated learning," arXiv preprint arXiv:2307.07477, 2023.
2. Z. Wang, G. Yang, H. Dai, et al., "DAFL: Domain adaptation-based federated learning for privacy-preserving biometric recognition," Future Generation Computer Systems, vol. 150, pp. 436-450, 2024.
3. X. Y. Liu, R. Zhu, D. Zha, et al., "Differentially private low-rank adaptation of large language model using federated learning," ACM Trans. Manage. Inf. Syst., vol. 16, no. 2, pp. 1-24, 2025.
4. W. Cui, "Vision-Oriented Multi-Object Tracking via Transformer-Based Temporal and Attention Modeling", Transactions on Computational and Scientific Methods, vol. 4, no. 11, 2024.
5. B. Fang and D. Gao, "Collaborative Multi-Agent Reinforcement Learning Approach for Elastic Cloud Resource Scaling", arXiv preprint arXiv:2507.00550, 2025.
6. T. Xu, X. Deng, X. Meng, H. Yang and Y. Wu, "Clinical NLP with Attention-Based Deep Learning for Multi-Disease Prediction", arXiv preprint arXiv:2507.01437, 2025.
7. Y. Kang, Y. He, J. Luo, et al., "Privacy-preserving federated adversarial domain adaptation over feature groups for interpretability," IEEE Trans. Big Data, vol. 10, no. 6, pp. 879-890, 2022.
8. D. Peterson, P. Kanani, V. J. Marathe, "Private federated learning with domain adaptation," arXiv preprint arXiv:1912.06733, 2019.
9. H. Li, X. Zhao, D. Guo, et al., "Federated domain-specific knowledge transfer on large language models using synthetic data," arXiv preprint arXiv:2405.14212, 2024.
10. Q. Wu, "Internal Knowledge Adaptation in LLMs with Consistency-Constrained Dynamic Routing", Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.
11. Y. Wang, "Structured Compression of Large Language Models with Sensitivity-aware Pruning Mechanisms", Journal of Computer Technology and Software, vol. 3, no. 9, 2024.
12. H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi and X. Yan, "Structuring Low-Rank Adaptation with Semantic Guidance for Model Fine-Tuning", 2025.
13. Y. Wu, Y. Qin, X. Su and Y. Lin, "Transformer-Based Risk Monitoring for Anti-Money Laundering with Transaction Graph Integration", 2025.

 doi:10.20944/preprints202509.1640.v1

14. Y. Xing, T. Yang, Y. Qi, M. Wei, Y. Cheng and H. Xin, "Structured Memory Mechanisms for Stable Context Representation in Large Language Models", arXiv preprint arXiv:2505.22921, 2025.
15. X. Quan, "Layer-Wise Structural Mapping for Efficient Domain Transfer in Language Model Distillation", Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.
16. Y. Sun, R. Zhang, R. Meng, L. Lian, H. Wang and X. Quan, "Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs", 2025.
17. H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan and Y. Xing, "Selective Knowledge Injection via Adapter Modules in Large - Scale Language Models", 2025.
18. L. Zhu, W. Cui, Y. Xing and Y. Wang, "Collaborative Optimization in Federated Recommendation: Integrating User Interests and Differential Privacy", Journal of Computer Technology and Software, vol. 3, no. 8, 2024.
19. M. Noble, A. Bellet, A. Dieuleveut, "Differentially private federated learning on heterogeneous data," Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, pp. 10110-10145, 2022.
20. C. Karouzos, G. Paraskevopoulos, A. Potamianos, "UDALM: Unsupervised domain adaptation through language modeling," arXiv preprint arXiv:2104.07078, 2021.
21. Y. Huang, D. Liu, Z. Zhong, et al., "KNN-Adapter: Efficient domain adaptation for black-box language models," arXiv preprint arXiv:2302.10879, 2023.
22. H. P. Wang, D. Chen, R. Kerkouche, et al., "Fedlap-dp: Federated learning by sharing differentially private loss approximations," arXiv preprint arXiv:2302.01068, 2023.