Structuring Low-Rank Adaptation with Semantic Guidance for Model Fine-Tuning

Hongye Zheng , Yumeng Ma , Yichen Wang , Guiran Liu , Zhen Qi , Xu Yan *

Posted Date: 18 June 2025

doi: 10.20944/preprints202506.1474.v1

Keywords: semantic perception fine-tuning; low-rank adaptation; question-answering system; efficient parameter learning

 doi:10.20944/preprints202506.1474.v1

Article

Structuring Low-Rank Adaptation with Semantic Guidance for Model Fine-Tuning

Hongye Zhen [1], Yumeng Ma [2], Yichen Wang [3], Guiran Liu [4], Zhen Qi [5] and Xu Yan [6,*]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Arizona State University Tempe, USA
[3] Georgia Institute of Technology, Atlanta, USA
[4] San Francisco State University, San Francisco, USA
[5] Northeastern University, Boston, USA
[6] Trine University, Phoenix, USA
* Correspondence: xyan232@my.trine.edu

Abstract: This paper addresses the challenges of fine-tuning efficiency and semantic adaptation in large language models for question answering tasks. It proposes a low-rank parameter adaptation method that incorporates semantic representations. While keeping the main model parameters frozen, the method introduces a semantic guidance function to improve traditional low-rank tuning strategies. This allows the parameter update process to dynamically align with input semantics, enhancing the model's ability to perceive complex semantic structures. The method embeds a semantic-aware module into the attention layers of the Transformer architecture. It uses representation vectors generated by a semantic encoder to guide the construction of low-rank matrices. In addition, a semantic similarity regularization term is applied to enforce consistency in the model's responses to semantically similar inputs. The method was evaluated across multiple experimental settings. These include comparisons with existing mainstream parameter-efficient fine-tuning approaches, analysis of adaptability to different QA types, and robustness under semantic perturbation. In all cases, the proposed method demonstrates strong accuracy, stability, and generalization ability. Furthermore, training loss curves show that the method achieves good convergence speed and training stability during optimization. Overall, the results indicate that the semantically guided low-rank adaptation strategy enhances the semantic understanding of QA systems while significantly reducing computational and storage costs during fine-tuning. This provides a simple yet robust solution for building efficient intelligent QA models.

Keywords: semantic perception fine-tuning; low-rank adaptation; question-answering system; efficient parameter learning

1. Introduction

In recent years, large language models (LLMs) have achieved breakthrough progress in natural language processing tasks. Leveraging their powerful capabilities in language understanding and generation, the performance of question answering (QA) systems has significantly improved. However, general-purpose LLMs are typically pre-trained on large-scale general corpora. When applied to domain-specific QA tasks such as in healthcare, law, or finance, they often face challenges such as insufficient knowledge transfer, semantic bias, and high deployment costs [1]. To address this issue, researchers have focused on fine-tuning LLMs for domain adaptation. This allows the models to incorporate domain knowledge and enhance their ability to understand questions. Yet, traditional full-parameter fine-tuning methods involve high computational and storage costs. They also suffer from parameter redundancy and unstable transfer performance, making it difficult to meet the demands of efficient, flexible, and controllable deployment [2].

 doi:10.20944/preprints202506.1474.v1

To alleviate these problems, parameter-efficient fine-tuning (PEFT) methods have recently emerged as a promising alternative for adapting LLMs in QA tasks. These methods introduce local adjustments to the model's parameter structure, such as inserting low-rank matrices or adding lightweight modules. By keeping the main model parameters frozen and updating only a small number of additional parameters, they significantly reduce training and storage costs. Among these approaches, low-rank adaptation techniques have gained wide attention due to their structural simplicity and scalability. However, current low-rank tuning methods often fall short in modeling semantic information. They typically perform linear adjustments in parameter space while ignoring the deeper connections between input semantics and internal model representations. As a result, they struggle to achieve robust generalization in complex QA scenarios [3].

QA tasks are essentially about semantic matching and generation. Their core lies in the model's accurate understanding of the semantic structure, intent expression, and background knowledge of the question. Therefore, combining semantic representation capabilities with parameter adaptation structures is key to improving fine-tuning performance. Introducing a semantically guided low-rank adaptation mechanism can align parameter updates more closely with input content and semantic needs [4,5]. This enables more targeted model adjustment. On this basis, the model can effectively achieve domain transfer and personalized QA tasks while maintaining the original knowledge structure. This approach is especially useful in scenarios with limited resources or strict inference cost requirements. It offers a better balance between accuracy and efficiency, enhancing the response quality and reliability of QA systems in real applications [6].

In addition, semantically aware low-rank tuning significantly improves a model's ability to adapt dynamically to varying input contexts, which is crucial for high-stakes applications in finance [7,8], healthcare [9,10], and data classification [11,12]. Unlike traditional static low-rank methods, this approach leverages semantic cues from the input to guide the fine-tuning process, enabling more precise and context-sensitive learning. This is particularly valuable in tasks such as fraud detection [13], clinical decision support [14], and personalized data interpretation [15,16], where understanding nuanced patterns can directly impact accuracy and trustworthiness. Parameter updates are no longer driven by uniform rules but are adjusted based on different semantic needs. This mechanism enhances the model's robustness to complex questions and supports higher-level QA tasks such as multi-turn dialogue and contextual reasoning. Especially in cross-domain and multi-style language settings, semantically guided adaptation helps maintain semantic consistency. It significantly improves user experience and the practical usability of the application.

2. Method

This study presents a low-rank adaptation method for fine-tuning large language models, explicitly incorporating semantic representations to improve semantic alignment and parameter efficiency. The method structurally enhances conventional low-rank tuning strategies by introducing a semantic guidance mechanism that modulates parameter updates based on the semantic characteristics of input data. This mechanism extends the low-rank update framework of Wang et al. [17] who demonstrated improvements in adaptation efficiency by optimizing the configuration of low-rank components. The proposed method maintains the parameters of the pre-trained model in a frozen state, consistent with the design used in parameter-efficient tuning. A semantic-aware low-rank adaptation module is embedded into the attention layers of the Transformer architecture, which serve as the primary locations for parameter injection. This design draws on the dynamic adaptation paradigm proposed by Cai, Kai, and Guo [18], who showed that low-rank modules can be made more adaptive to input variance. Semantic vectors, derived from a lightweight semantic encoder, guide the construction and updating of the low-rank matrices. These vectors control both the direction and magnitude of injected updates, enabling more precise and semantically consistent parameter modulation. The semantic integration approach aligns with findings by Xu et al. [19], who emphasized the utility of semantically structured representations in enhancing Transformer-based models.

 doi:10.20944/preprints202506.1474.v1

The model architecture, including the semantic-aware adaptation mechanism, is illustrated in Figure 1.

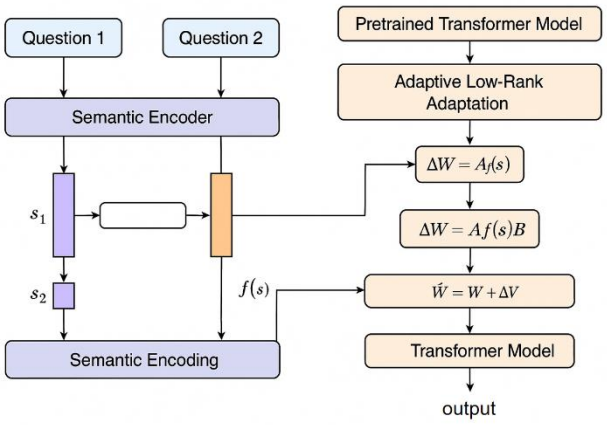

Figure 1. Overall model architecture diagram.

In terms of method modeling, firstly, a weight matrix in the pre-trained language model is set to $W \in R^{d_{out} \times d_{in}}$, and the traditional low-rank adaptation method replaces it with:

$$\hat{W} = W + \Delta W = W + AB$$

Among them are $A \in R^{d_{out} \times r}$, $B \in R^{r \times d_m}$, and $r << \min(d_{out}, d_{in})$. In order to introduce semantic guidance, we introduce a semantic representation vector $s \in R^{d_s}$ and construct a semantic regulation function $f_s : R^{d_s} \to R^{r \times r}$ to dynamically adjust the low-rank structure:

$$\Delta W = A f_s(s) B$$

The function f_s can be expressed as a lightweight feedforward network, which is used to perform linear transformation and nonlinear activation on the semantic vector, generate a semantic weight matrix, and achieve alignment between the semantics and parameter update path.

A semantic distance constraint mechanism is introduced to enhance the semantic adaptability of the fine-tuning process. This constraint enforces that parameter updates resulting from semantically similar inputs remain consistent, thereby stabilizing the model's behavior across related queries. The mechanism is informed by prior work demonstrating that semantic similarity constraints can regulate transformation consistency in language model outputs, particularly when applied to structured semantic tasks [20]. Furthermore, research on hierarchical semantic representations supports the design of constraints that reflect deeper relational structures, rather than relying solely on surface-level similarity [21]. This constraint is formulated as an additional loss term and is applied during training to limit divergence in parameter space between semantically related inputs. The result is a more semantically coherent and generalizable adaptation strategy. Specifically, assuming that the semantic representations of the two input sentences are s_1 and s_2, and their corresponding parameter injection items are ΔW_1 and ΔW_2, the following regularization term is introduced to control the semantic consistency:

$$L_{sem} = \| \Delta W_1 - \Delta W_1 \|_F^2 \cdot sim(s_1, s_2)$$

Where $sim(s_1, s_2)$ represents semantic similarity (such as cosine similarity). This regularization term encourages the parameter injection of semantically similar inputs to have

 doi:10.20944/preprints202506.1474.v1

structural similarity, thereby improving the smoothness and generalization ability of the model in the semantic space.

In terms of the overall training goal, the model jointly optimizes the original task loss of the language model and the semantic constraint loss to form a joint loss function:

$$L_{total} = L_{task} + \lambda L_{sem}$$

Where λ is a hyperparameter that adjusts the contribution of the semantic regularization term to the total loss. Through the above method, the model can still dynamically adapt according to the input semantic content while adjusting only a few parameters, realizing an efficient and precision-controlled fine-tuning solution for the question-answering system. While maintaining low computational overhead, the entire framework significantly enhances the model's semantic perception and expression capabilities, making it particularly suitable for application in semantically complex and diverse question-answering tasks.

3. Experimental Results

3.1. Dataset

This study uses HotpotQA as the primary dataset to build and evaluate the fine-tuning performance of large language models on multi-hop question answering tasks. HotpotQA is a challenging open-domain QA dataset containing approximately 113,000 question-answer pairs. Each question requires the model to perform multi-hop reasoning across multiple documents to arrive at the correct answer. This characteristic makes it well-suited for evaluating semantically guided fine-tuning methods, especially when dealing with complex reasoning paths and long document contexts.

The data in HotpotQA is constructed from Wikipedia. Each sample includes a natural language question, multiple related passages (including gold-standard supporting sentences), and a clearly defined answer. Unlike traditional single-hop QA tasks, this dataset emphasizes the need to "find and combine multiple pieces of evidence" to answer each question. This not only increases the difficulty of the task but also raises the demands on semantic modeling and cross-paragraph information alignment. Therefore, HotpotQA is highly representative in terms of semantic awareness, information integration, and knowledge reasoning.

In addition, HotpotQA provides labels that distinguish between "bridge" and "comparison" questions, further increasing the diversity of QA types. Bridge questions require the model to connect multiple entities for reasoning across different contexts. Comparison questions involve understanding quantitative relations, temporal order, or contrast between entities. This variety of question types helps evaluate the stability and generalization ability of the proposed low-rank semantic adaptation mechanism under different semantic structures.

3.2. Experimental Results

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

Table 1. Comparative experimental results.

Method	EM(%)	F1(%)	SemSim
T5-Base [22]	67.2	81.0	0.784
Adapater-T5 [23]	68.5	82.1	0.791
LoRA-T5 [24]	70.3	83.5	0.806
Prompt-Tuning [25]	65.4	78.7	0.763
Ours	72.1	85.0	0.829

 doi:10.20944/preprints202506.1474.v1

Experimental results on the HotpotQA dataset demonstrate that the proposed semantically guided low-rank adaptation method (SeLoRA) outperforms mainstream parameter-efficient fine-tuning approaches. SeLoRA achieves 72.1% in Exact Match (EM) and 85.0% in F1 score, marking improvements of 1.8 and 1.5 percentage points, respectively, over standard LoRA-T5. These gains highlight the effectiveness of semantic guidance in enhancing model comprehension and generation while keeping core parameters frozen. Compared to traditional full-parameter tuning (T5-Base), which achieves a lower SemSim score of 0.784, and prompt-based methods that lack structural and semantic integration, SeLoRA delivers superior semantic consistency and task performance. While methods like Adapter-T5 and LoRA-T5 show stable results, they fall short in semantic alignment due to the absence of explicit semantic control. By embedding semantic representations into the low-rank update process, SeLoRA enables more informed parameter adjustments, leading to robust performance across complex reasoning tasks. The SemSim metric, which measures semantic coherence between questions and answers, confirms SeLoRA's advantage in deep reasoning and alignment. The impact of fine-tuning across different question-answering types is further illustrated in Figure 2.

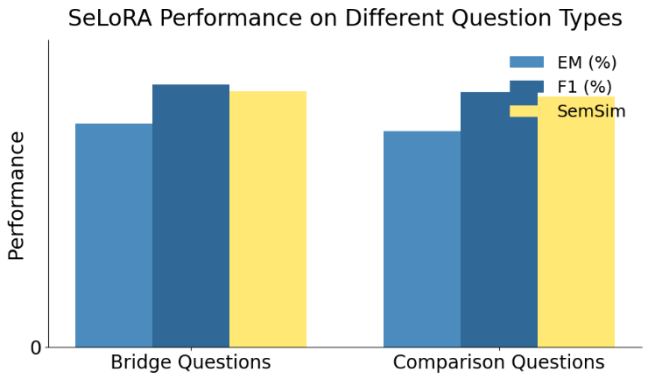

Figure 2. Evaluation of fine-tuning effects under different question-answering types.

The figure shows that the proposed algorithm achieves stable and strong performance across different types of question answering tasks, especially in terms of F1 score and semantic similarity (SemSim). For bridge questions, the model reaches a high F1 score, indicating strong abilities in information integration and cross-paragraph reasoning. This performance suggests that the low-rank adaptation with semantic guidance effectively enhances the model's perception and processing of multi-hop information chains, improving its ability to model complex semantic relationships.

In terms of Exact Match (EM), the model performs slightly better on comparison questions. This suggests that the proposed algorithm can generate more precise answers when dealing with questions that have clear structure and explicit semantic contrast. Comparison questions often involve reasoning about quantities, time, or entity differences. Compared to bridge questions, they rely less on textual connections and benefit more from semantically controlled parameter paths.

The results on semantic similarity (SemSim) further confirm the model's consistency and robustness across different semantic tasks. Whether integrating clues across documents or comparing concepts, the model's outputs maintain a high level of semantic alignment with the reference answers. This shows that the semantically aware parameter adjustment process enables the model to better capture deep contextual meaning, reducing semantic drift and comprehension errors.

In conclusion, the proposed fine-tuning method shows strong adaptability and semantic modeling capability in both types of QA tasks. The results demonstrate the generality and effectiveness of the semantically guided low-rank adaptation mechanism across diverse QA scenarios. The model performs well not only in surface-level accuracy but also in fine-grained semantic understanding. This provides both theoretical foundation and practical support for high-quality deployment of complex QA systems.

 doi:10.20944/preprints202506.1474.v1

This paper also presents an adversarial experiment on the robustness of the model under semantic perturbations, and the experimental results are shown in Figure 3.

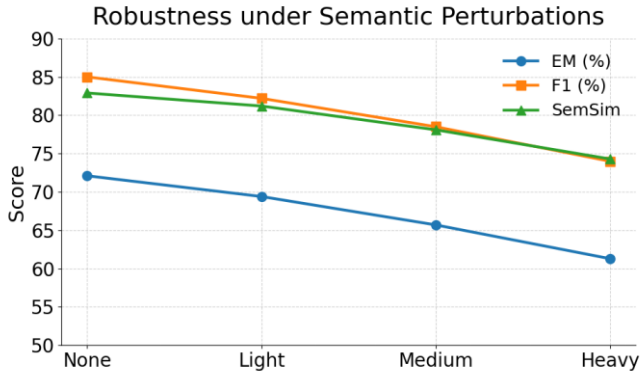

Figure 3. Adversarial Experiments on Model Robustness Under Semantic Perturbations.

Figure 4 shows that as the level of semantic perturbation increases, the model's performance on all metrics progressively declines, indicating a degree of vulnerability to input variation. While F1 and SemSim scores exhibit only moderate decreases, the sharper drop in Exact Match (EM) suggests that surface-level alterations more readily disrupt precise outputs. The model performs best with unperturbed inputs, demonstrating effective semantic modeling under standard conditions. However, under heavy perturbation, the EM score falls to 61.3%, revealing limitations in the robustness of the current semantic guidance mechanism. These results highlight the need for enhanced strategies, such as adversarial training or context consistency techniques, to improve the model's adaptability to semantic variation.

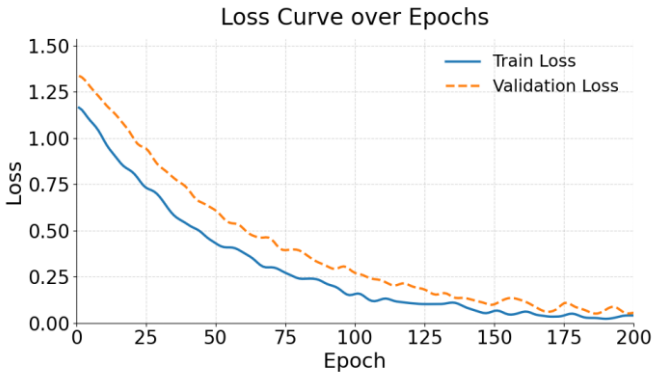

Figure 4. Loss Curve over Epochs.

The figure shows that the proposed fine-tuning algorithm demonstrates good convergence during training. Both the training loss and validation loss drop rapidly in the early stages and stabilize in the later stages. This indicates that the model completes the main semantic adaptation and parameter convergence within a short training period. It suggests that the semantically guided low-rank adaptation mechanism can quickly capture the semantic structure of input data with limited parameter updates.

Further observation shows that the validation loss remains close to the training loss throughout the training process. No significant fluctuations or signs of overfitting are observed. This suggests that the method maintains strong learning ability while achieving good generalization performance. Especially during the later stages of training, when the loss is low, the validation loss remains stable.

 doi:10.20944/preprints202506.1474.v1

This confirms the model's improved semantic understanding on unseen data and reflects the effectiveness of the semantic control mechanism in ensuring robustness.

Overall, the loss curves validate the practicality and convergence efficiency of the proposed method in parameter-efficient fine-tuning tasks. The model not only quickly absorbs domain-specific semantic features in the early phase but also maintains a stable optimization trajectory throughout the training process. This provides a solid foundation for deploying high-performance and low-cost models in complex QA tasks.

4. Conclusions

This study focuses on parameter-efficient fine-tuning of large language models for question answering tasks. It proposes a low-rank adaptation method that integrates semantic representations. By introducing a semantic guidance mechanism into the low-rank matrix structure, the method aligns parameter updates with input semantics during fine-tuning. This enhances semantic modeling and QA performance while maintaining a lightweight model design. Compared to traditional full-parameter tuning and structure-based injection methods, the proposed approach achieves better results across multiple metrics. This confirms its effectiveness and adaptability in complex semantic tasks.

The method was evaluated across various experimental dimensions, including model performance comparison, analysis of QA types, robustness under semantic perturbations, and training loss convergence. In all settings, the method shows strong generalization and training stability. It performs especially well in multi-hop reasoning and semantically diverse QA tasks. The semantic-aware mechanism helps the model better understand complex contextual structures and generate accurate responses. These results provide method-level support for improving the controllability, adaptability, and deployment efficiency of large models in practical applications.

At the application level, the proposed method has direct value for domains requiring high semantic precision, such as medical, legal, and financial question answering. In these scenarios, user queries often involve deep reasoning, technical term understanding, and contextual consistency. Traditional methods struggle to balance performance with training cost. The findings of this study offer a general framework for customized fine-tuning under low-resource conditions. This enhances the practical value of models in professional contexts and supports the development and deployment of high-quality intelligent QA systems. Future work may further expand the generalization of the semantic guidance mechanism. This includes integrating retrieval-augmented modules, multimodal semantic injection, or cross-lingual semantic alignment to handle more complex input structures. Research can also explore how to combine semantic control strategies with task planning, long-context modeling, and dialogue memory retention. Combining semantic awareness with other parameter-efficient fine-tuning techniques may advance large language models toward higher levels of intelligence in QA systems, virtual assistants, and human-computer interaction.

References

1. D. Ding, Y. Qin, G. Yang, et al., “Parameter-efficient fine-tuning of large-scale pre-trained language models,” Nature Machine Intelligence, vol. 5, no. 3, pp. 220-235, 2023.
2. R. Xu, F. Luo, Z. Zhang, et al., “Raise a child in large language model: Towards effective and generalizable fine-tuning,” arXiv preprint arXiv:2109.05687, 2021.
3. Y. Liu, A. Singh, C. D. Freeman, et al., “Improving large language model fine-tuning for solving math problems,” arXiv preprint arXiv:2310.10047, 2023.
4. S. Zhang, L. Dong, X. Li, et al., “Instruction tuning for large language models: A survey,” arXiv preprint arXiv:2308.10792, 2023.
5. B. Gunel, J. Du, A. Conneau, et al., “Supervised contrastive learning for pre-trained language model fine-tuning,” arXiv preprint arXiv:2011.01403, 2020.

 doi:10.20944/preprints202506.1474.v1

6. S. Höglund and J. Khedri, "Comparison between RLHF and RLAIF in fine-tuning a large language model," 2023.
7. J. Liu, X. Gu, H. Feng, Z. Yang, Q. Bao and Z. Xu, "Market Turbulence Prediction and Risk Control with Improved A3C Reinforcement Learning," arXiv preprint arXiv:2503.02124, 2025.
8. X. Du, "Financial Text Analysis Using 1D-CNN: Risk Classification and Auditing Support," arXiv preprint arXiv:2503.02124, 2025.
9. Y. Xiang, Q. He, T. Xu, R. Hao, J. Hu and H. Zhang, "Adaptive Transformer Attention and Multi-Scale Fusion for Spine 3D Segmentation," arXiv preprint arXiv:2503.12853, 2025.
10. T. Xu, Y. Xiang, J. Du and H. Zhang, "Cross-Scale Attention and Multi-Layer Feature Fusion YOLOv8 for Skin Disease Target Detection in Medical Images," Journal of Computer Technology and Software, vol. 4, no. 2, 2025.
11. Y. Deng, "A Hybrid Network Congestion Prediction Method Integrating Association Rules and LSTM for Enhanced Spatiotemporal Forecasting," Transactions on Computational and Scientific Methods, vol. 5, no. 2, 2025.
12. M. Li, R. Hao, S. Shi, Z. Yu, Q. He and J. Zhan, "A CNN-Transformer Approach for Image-Text Multimodal Classification with Cross-Modal Feature Fusion," arXiv preprint arXiv:2504.00282, 2025.
13. J. Gong, Y. Wang, W. Xu and Y. Zhang, "A Deep Fusion Framework for Financial Fraud Detection and Early Warning Based on Large Language Models," Journal of Computer Science and Software Applications, vol. 4, no. 8, 2024.
14. R. Hao, Y. Xiang, J. Du, Q. He, J. Hu and T. Xu, "A Hybrid CNN-Transformer Model for Heart Disease Prediction Using Life History Data," arXiv preprint arXiv:2503.02124, 2025.
15. X. Wang, "Medical Entity-Driven Analysis of Insurance Claims Using a Multimodal Transformer Model," Journal of Computer Technology and Software, vol. 4, no. 3, 2025.
16. Y. Zhang, J. Liu, J. Wang, L. Dai, F. Guo and G. Cai, "Federated Learning for Cross-Domain Data Privacy: A Distributed Approach to Secure Collaboration," arXiv preprint arXiv:2504.00282, 2025.
17. Y. Wang, Z. Fang, Y. Deng, L. Zhu, Y. Duan and Y. Peng, "Revisiting LoRA: A Smarter Low-Rank Approach for Efficient Model Adaptation," arXiv preprint arXiv:2505.0618, 2025.
18. G. Cai, A. Kai and F. Guo, "Dynamic and Low-Rank Fine-Tuning of Large Language Models for Robust Few-Shot Learning," Transactions on Computational and Scientific Methods, vol. 5, no. 4, 2025.
19. Z. Xu, Y. Sheng, Q. Bao, X. Du, X. Guo and Z. Liu, "BERT-Based Automatic Audit Report Generation and Compliance Analysis," arXiv preprint arXiv:2505.0618, 2025.
20. R. Wang, "Joint Semantic Detection and Dissemination Control of Phishing Attacks on Social Media via LLama-Based Modeling," arXiv preprint arXiv:2504.00282, 2025.
21. G. Cai, J. Gong, J. Du, H. Liu and A. Kai, "Investigating Hierarchical Term Relationships in Large Language Models," Journal of Computer Science and Software Applications, vol. 5, no. 4, 2025.
22. E. Lehman and A. Johnson, "Clinical-t5: Large language models built using mimic clinical text," PhysioNet, 2023.
23. Z. Hu, L. Wang, Y. Lan, et al., "LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models," arXiv preprint arXiv:2304.01933, 2023.
24. E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-rank adaptation of large language models," ICLR, vol. 1, no. 2, pp. 3, 2022.
25. C. Peng, X. I. Yang, K. E. Smith, et al., "Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction," Journal of Biomedical Informatics, vol. 153, pp. 104630, 2024.