Article

Not peer-reviewed version

Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models

Hongye Zheng , Lipeng Zhu , Wanyu Cui , Ray Pan , Xu Yan , Yue Xing *

Posted Date: 3 June 2025

doi: 10.20944/preprints202506.0147.v1

Keywords: dynamic fine-tuning; knowledge fusion; gating mechanism; multi-hop reasoning

Article

Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models

Hongye Zheng [1], Lipeng Zhu [2], Wanyu Cui [3], Ray Pan [4], Xu Yan [5] and Yue Xing [6,*]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Johns Hopkins University, Baltimore, USA
[3] University of Southern California, Los Angeles, USA
[4] Independent Researcher, Seattle, USA
[5] Trine University, Phoenix, USA
[6] University of Pennsylvania, Philadelphia, USA
* Correspondence: jasmineyxing@gmail.com

Abstract: This paper addresses key challenges in knowledge injection for large language models, including static representation, difficulty in updating, and limited domain adaptability. It proposes a dynamic fine-tuning framework designed for knowledge injection. The framework is based on parameter-efficient tuning strategies and introduces learnable adapter modules and gating mechanisms. These components enable selective integration and dynamic control of both structured and unstructured external knowledge. In the model design, a query encoder extracts semantic vectors from the input text. These vectors are matched with an external knowledge base to construct a dynamic knowledge subset. This subset guides task generation. Adapter modules and gating units are then applied across model layers to adjust knowledge enhancement. This ensures that external knowledge contributes to reasoning while preserving the model's original language capabilities. A unified joint loss function is also introduced. It coordinates the optimization between language modeling and knowledge alignment objectives. To evaluate the proposed method, the WikiHop dataset for multi-hop question answering is used. The model's behavior is analyzed under various experimental settings, including different parameter update ratios, knowledge densities, and cross-domain transfer scenarios. The results show that the method achieves strong performance on key metrics such as knowledge recall, F1 score, and inference efficiency, even with low parameter update ratios. This demonstrates the practicality and stability of the approach in dynamic knowledge integration tasks. The proposed work offers a flexible and efficient technical path for knowledge injection and domain adaptation in large language models.

Keywords: dynamic fine-tuning; knowledge fusion; gating mechanism; multi-hop reasoning

I. Introduction

With the rapid development of artificial intelligence, large language models (LLMs) have demonstrated wide applicability across various natural language processing tasks due to their powerful capabilities in language understanding and generation[1]. However, as model sizes continue to grow and application scenarios diversify, an urgent challenge arises: how to efficiently inject domain-specific knowledge into pretrained models to achieve more accurate and reliable performance in specific tasks[2]. The traditional pretraining-finetuning paradigm struggles with tasks requiring frequent knowledge updates or domain expertise. It faces challenges such as knowledge lag, lack of interpretability, and reduced generalization. A more flexible and efficient mechanism for knowledge injection is urgently needed.

In practice, many applications demand highly specialized knowledge from language models[3]. Fields such as healthcare, human-computer interaction, and finance not only require basic language understanding but also necessitate mastery of complex domain-specific knowledge [4–7]. Although existing pretrained models are trained on general corpora and exhibit strong language modeling capabilities, their performance often suffers in specific domains due to missing or inaccurate knowledge. Moreover, the knowledge embedded in pretrained models tends to be static. Once training is complete, updating or adjusting this knowledge becomes difficult. This is particularly unsuitable for scenarios where knowledge evolves rapidly. Therefore, a critical challenge is to achieve dynamic injection and updating of domain knowledge while preserving the model's general capabilities[8].

Knowledge injection serves as a vital approach to bridge language models with external knowledge, significantly enhancing knowledge density and task adaptability[9]. As LLMs scale up rapidly, there is a growing research interest in integrating structured or unstructured knowledge into models in a low-cost and minimally intrusive manner. Traditional methods rely heavily on static fusion or full-parameter finetuning, which are costly, inefficient, and ill-suited for frequently updated knowledge. Recent advances in parameter-efficient finetuning methods, such as LoRA and Adapters, along with retrieval-augmented generation mechanisms, offer new directions for dynamic knowledge injection. However, challenges remain in terms of consistency, stability, and scalability[10].

Against this backdrop, building a dynamic finetuning framework for LLMs tailored to knowledge injection tasks becomes particularly important. Such a framework should balance the completeness of knowledge acquisition and the accuracy of model generation. It should also support incremental updates, selective injection, and dynamic control during inference[11]. By designing a flexible and scalable finetuning mechanism along with an effective knowledge management strategy, the framework can enhance the model's domain knowledge comprehension and application without compromising its original language capabilities. This not only facilitates real-world deployment of LLMs but also lays a foundation for constructing sustainable and evolving intelligent systems.

Overall, research on dynamic finetuning frameworks for knowledge injection in LLMs holds significant theoretical and practical value. It promotes deep integration between LLMs and knowledge engineering and supports the development of controllable, interpretable, and scalable intelligent systems. As intelligent applications demand higher knowledge density and professional competence, building an efficient, stable, and continuously updatable knowledge injection framework will play an increasingly central role in key scenarios such as question answering, dialogue systems, automatic summarization, and legal decision support [12–14]. Thus, exploring dynamic finetuning mechanisms aligns with the frontier of AI development and contributes to the advancement of knowledge-enhanced intelligent systems.

II. Method

This work introduces a dynamic fine-tuning approach for large language models, designed to seamlessly integrate external domain knowledge while retaining the model’s core language understanding and generation abilities. Building on the parameter-efficient design proposed by Deng [15], we move away from full-model updates and instead adopt lightweight, learnable modules that can be locally tuned. This modular approach allows the model to adapt incrementally to new knowledge, offering both flexibility and efficiency.

Structured and unstructured knowledge is explicitly encoded and injected through targeted mechanisms. Drawing from the insights of Wu et al. [16], we use enhanced embedding strategies that improve the model’s sensitivity to domain-specific entities and relationships. These strategies help ensure that external knowledge is not simply added to the input but actively guides the model’s internal reasoning processes. To further refine how knowledge is applied, our method employs a dynamic injection mechanism inspired by Yu et al. [17], where the integration of external knowledge is controlled through adaptive gating. These gates selectively activate relevant information, enabling the model to respond more precisely to task demands without overwhelming its core language representations. Together, these components form an architecture that supports efficient knowledge injection, fine-grained control, and reduced overhead. As shown in Figure 1, the model incorporates learnable modules across key layers, enabling targeted updates and smooth integration of external knowledge as needed.

First, in order to achieve structured representation and efficient retrieval of knowledge, a knowledge base $K = \{k_1, k_2, ..., k_n\}$ is set, where each $X = \{x_1, x_2, ..., x_m\}$ represents a knowledge fragment or triple. For the input text sequence $Eq(\cdot)$, it is mapped to a query vector $q = Eq(X)$ through a query encoder D. In the process of knowledge injection, relevant knowledge items are selected by vector similarity calculation:

$$sim(q, k_i) = \frac{q \cdot E_k(k_i)}{\| q \| \| E_k(k_i) \|}$$

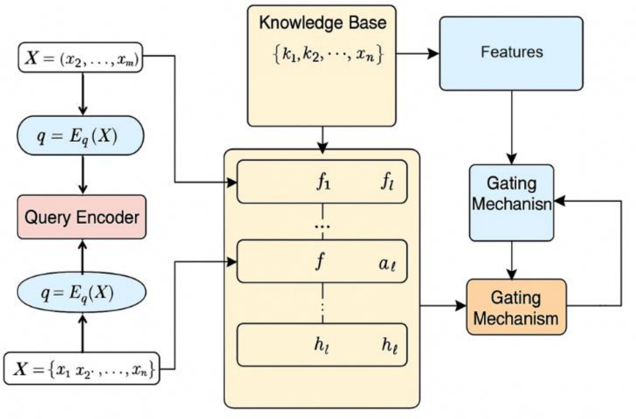

Figure 1. Overall model architecture diagram.

$E_k(\cdot)$ is the encoder function of the knowledge item, and finally the first T pieces of relevant knowledge are selected according to the similarity to form the dynamic knowledge set $K' \subset K$.

Secondly, in order to achieve local regulation and knowledge fusion of the model, this method introduces a lightweight learnable adapter module A_θ to enhance the knowledge of the original model layer output h_l. Define the output of each layer as $h_l = f_l(h_{l-1})$, and the role of the adapter module is:

$$\widetilde{h}_l = h_l + A_\theta(h_l, K')$$

$A_\theta(\cdot)$ receives the current layer output and knowledge subset as input, introduces knowledge enhancement information through a feedforward structure, and keeps the overall structure of the model unchanged. This design allows local and controllable knowledge injection, thus avoiding interference with the global behavior of the model.

To dynamically regulate how external knowledge influences the generation process, a knowledge gating mechanism is integrated into the model. This mechanism applies learned, context-dependent weights to the outputs of each layer, ensuring that knowledge is neither excessively applied nor underutilized. The gates respond to both the semantic content of the input and the relevance of the retrieved knowledge, enabling the model to dynamically adjust its behavior as generation progresses.

This adaptive control plays a critical role in aligning external knowledge with the model's internal reasoning pathways. It draws on insights from recent work showing that hierarchical and task-specific modulation can significantly enhance the model's ability to incorporate layered or domain-sensitive knowledge [18], particularly in settings requiring precision and interpretability [19]. The gating function is formally defined as follows:

$$g_l = \sigma(W_g[h_l;\tau_l] + b_g)$$

$$h_l = g_l \otimes \tau_l + (1 - g_l) \otimes h_l$$

$\tau_l = A_\theta(h_l, K'), \sigma(\cdot)$ is the Sigmoid function, $[h_l;\tau_l]$ represents vector concatenation, and g_l controls the strength of knowledge fusion. This mechanism enables the model to selectively absorb or ignore knowledge information in different contexts, improving the flexibility and robustness of knowledge injection.

Finally, to ensure the consistency and traceability of knowledge during task fine-tuning, a joint optimization objective function is designed to jointly train the language modeling loss L_{LM} and the knowledge alignment loss L_{KN}:

$$L = L_{LM} + \lambda L_{KN}$$

λ is a regulating factor, and L_{KN} is used to constrain the consistency between the model generation results and the dynamic knowledge set, such as by minimizing the KL divergence between the generated probability distribution and the knowledge target distribution. This joint training strategy ensures the coordinated optimization of language modeling and knowledge injection goals, thereby achieving robust adaptation and scalable migration of knowledge-enhanced language models in dynamic environments.

III. Experiment

A. Datasets

This study utilizes the WikiHop dataset as the primary evaluation corpus for knowledge injection tasks. WikiHop is a multi-hop reading comprehension dataset constructed from the hyperlink structure of Wikipedia. It is designed to assess how models reason across multiple documents to find answers. Each sample comprises a query question, multiple supporting passages, and a set of candidate answers. The model must select the correct answer based on the supporting evidence. This multi-hop question-answering format inherently relies on external knowledge. It serves as an ideal experimental platform for testing knowledge enhancement mechanisms.

The dataset spans a wide range of topics, including geography, history, politics, and culture. Its data structure supports testing a model's ability to integrate information across passages and entities. Questions in WikiHop are often not answerable from a single document. Models typically need to combine information from multiple sources to draw conclusions. This makes the dataset highly relevant and challenging for evaluating knowledge injection and dynamic finetuning strategies.

In addition, WikiHop provides clear data splits, including training, validation, and test sets. This facilitates model development and tuning. Its standardized format and public benchmarks have made it widely used in research on multi-hop reasoning and knowledge-based question answering. It offers a stable and reliable foundation for evaluating the dynamic knowledge finetuning framework proposed in this study.

B. Experimental Results

This paper first gives the comparative experimental results, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	Knowledge	F1 Score(%)	Inference

	Recall(%)		Time(ms)
T5-base + Full Fine-tuning[20]	68.4	55.1	82
T5-base + Adapter Tuning[21]	71.3	60.5	76
T5-base + Prefix Tuning[22]	73.6	63.2	79
T5-base + LoRA[23]	70.8	58.7	88
T5-base+Ours	75.4	67.9	81

The results in Table 1 indicate that full-parameter fine-tuning of T5-Base yields modest knowledge recall (68.4 %) and F1 (55.1 %), despite acceptable inference latency; however, the need to update all weights incurs high training overhead and precludes on-the-fly knowledge updates. By contrast, parameter-efficient methods—Adapter Tuning and Prefix Tuning—achieve higher recall (73.6 % for Prefix) and F1 (63.2 %) at minimal cost, with Prefix Tuning's input-level guidance markedly enhancing knowledge integration, though still lacking fine-grained control over injection. LoRA, which perturbs low-rank weight subsets, outperforms full fine-tuning but trails adapter-based approaches, underscoring that structured control, not mere parameter compression, drives multi-hop QA performance.

Our proposed dynamic fine-tuning framework—combining adaptive adapters with a gating mechanism—delivers the best results (75.4 % recall, 67.9 % F1) while maintaining a lightweight 81 ms inference time. This demonstrates its ability to selectively inject and synergize knowledge representations without significant computational burden. Figure 2 further analyzes how varying the ratio of updated parameters affects task performance, corroborating the framework's efficiency and practicality for scalable, flexible knowledge injection.

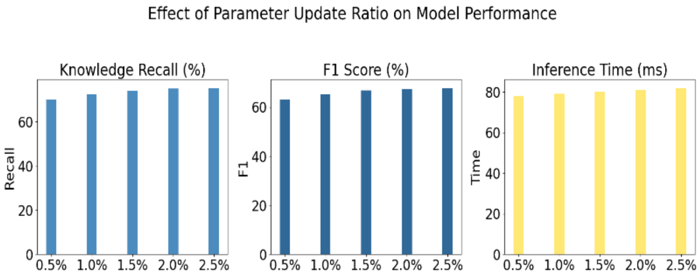

Figure 2. Experiment on the relationship between model parameter update ratio and task performance.

As shown in the results of Figure 2, when the proportion of model parameters being updated increases from 0.5% to 2.5%, both knowledge recall and F1 score show a stable upward trend. This indicates that moderately increasing parameter participation has a positive impact on the effectiveness of knowledge injection. It reflects the sensitivity of the dynamic finetuning mechanism to knowledge integration. Even with minimal parameter intervention, the model can still effectively incorporate external knowledge. Specifically, when the update ratio is between 2.0% and 2.5%, the improvement in performance becomes less pronounced. This suggests diminishing marginal returns for knowledge injection. It also implies that large-scale modification of pretrained parameters is not necessary. A small number of adapters or gating units can achieve significant knowledge enhancement. This reduces training costs and helps avoid overfitting.

In terms of inference efficiency, although the proportion of updated parameters increases, inference time remains stable, ranging from 78 ms to 82 ms. This shows that the designed finetuning structure has minimal impact on computational resources during deployment. Compared to traditional full-parameter finetuning, the proposed method is more practical, especially in scenarios with limited resources or frequent knowledge updates.

Taken together, these results confirm that the proposed dynamic finetuning framework can enhance knowledge capabilities through minor parameter adjustments without compromising inference efficiency. This demonstrates both the efficiency and stability of the approach. It also provides strong empirical support for future work on more fine-grained knowledge control mechanisms. Furthermore, this paper also presents an analysis of the model response behavior based on changes in knowledge density, and the experimental results are shown in Figure 3.

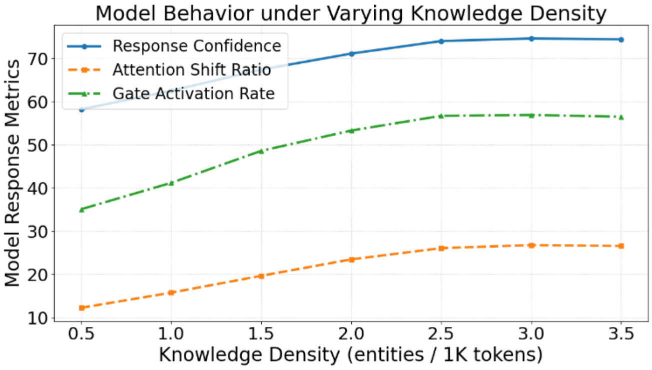

Figure 3. Analysis of model response behavior based on knowledge density changes.

As shown in Figure 3, with the gradual increase in knowledge density, the model's response confidence exhibits a stable upward trend and reaches saturation when the density approaches 3.0. This indicates that the proposed dynamic finetuning mechanism has strong knowledge integration capabilities. It can effectively activate the internal knowledge representation of the language model in dense knowledge scenarios, enhancing the confidence and consistency of responses.

The attention shift ratio increases as knowledge density rises. This shows that during inference, the model more frequently shifts attention from the original context to knowledge-relevant segments. This shift suggests that the model has a certain degree of structured knowledge awareness. It can flexibly redirect reasoning from surface linguistic cues to high-value knowledge clues. This is a key manifestation of effective dynamic knowledge injection. The gate activation rate also increases significantly with higher knowledge density and tends to stabilize when the density exceeds 2.5. This demonstrates that the gating mechanism can adapt to varying knowledge loads. As the amount of external knowledge grows, the model selectively strengthens the knowledge pathway. It retains linguistic information while guiding the model to rely more on knowledge-enhanced channels. Overall, the experiment confirms the effectiveness of the proposed dynamic injection framework under different knowledge distribution conditions. The model can adjust its internal processing strategy in response to knowledge density. Through the use of gating structures and attention reallocation, it achieves efficient knowledge scheduling. This provides strong empirical support for controllable knowledge modeling in complex language tasks.

This paper also presents a test on the impact of different domain knowledge on the language model transfer capability, and the experimental results are shown in Figure 4.

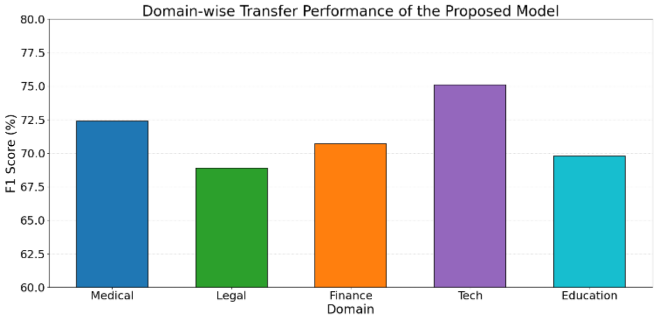

Figure 4. Test of the impact of different domain knowledge on language model transfer ability.

As shown in the results of Figure 4, the proposed dynamic finetuning model shows varying levels of knowledge transfer ability across different domains. Its performance is most prominent in the technology domain, achieving an F1 score of 75.1%. This suggests that in domains with dense structured information and stable terminology distribution, the dynamic knowledge injection mechanism can efficiently capture and integrate relevant knowledge. It enables strong cross-task generalization.

In contrast, the transfer performance in the legal and education domains is slightly lower, with F1 scores of 68.9% and 69.8% respectively. This may be due to the more complex language styles and less structured knowledge organization in these fields. As a result, the model faces greater challenges in both knowledge extraction and generation. These domains often rely on implicit reasoning and unstructured background knowledge, which increases the demands on the adaptability of the dynamic injection mechanism.

The financial and medical domains show relatively stable performance. This indicates that in contexts with clear term boundaries and hierarchical knowledge structures, the dynamic finetuning strategy can maintain knowledge consistency and generation quality. Overall, the experiment demonstrates both the flexibility and limitations of the dynamic knowledge injection framework in cross-domain transfer. The model performs well in domains with clear knowledge structures and reasoning paths. However, in tasks with high semantic complexity or weak knowledge organization, there remains a need to optimize knowledge representation and injection strategies to improve domain generalization.

IV. Conclusion

This study proposes a dynamic fine-tuning framework for knowledge injection in large language models. It enables efficient integration and updating of knowledge while preserving the model's general language capabilities. The method introduces lightweight, learnable modules and gating mechanisms. These allow the model to dynamically adjust the strength of knowledge injection during inference. The approach significantly improves performance on complex tasks such as multi-hop question answering and cross-domain reasoning. Through flexible design and efficient parameter updates, the framework achieves a good balance among training cost, knowledge adaptability, and inference efficiency.

A series of experiments validate the advantages of the proposed method under varying knowledge densities, domain transfer conditions, and parameter sensitivity settings. The results demonstrate the synergy between knowledge expression and language understanding enabled by dynamic injection. Compared with traditional static finetuning methods, the dynamic framework better meets real-world demands for rapid knowledge evolution and frequent task switching. It offers a new technical path to address the challenges of knowledge lag and transfer bottlenecks in pretrained models. This capability is not only of academic value but also practically significant for building domain-specific intelligent systems such as question answering, legal assistance, and medical reasoning.

As AI models are increasingly applied in real-world scenarios, there is growing demand for scalability, interpretability, and controllability. The dynamic knowledge injection mechanism provides finer control channels, allowing the model to flexibly select knowledge and adjust strategies in complex environments. The findings of this study may promote deeper integration of pretrained language models with knowledge graphs and external retrieval systems. This could facilitate more efficient deployment of natural language processing technologies in key industries.As technology continues to evolve, the dynamic knowledge injection framework may become a vital pillar in advancing the next generation of intelligent language systems.

References

1. T. Susnjak et al., "Automating research synthesis with domain-specific large language model fine-tuning," ACM Transactions on Knowledge Discovery from Data, vol. 19, no. 3, pp. 1–39, 2025.
2. R. Pan et al., "LISA: layerwise importance sampling for memory-efficient large language model fine-tuning," Advances in Neural Information Processing Systems, vol. 37, pp. 57018–57049, 2024.
3. 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.
4. J. Zhan, "Single-Device Human Activity Recognition Based on Spatiotemporal Feature Learning Networks," Transactions on Computational and Scientific Methods, vol. 5, no. 3, 2025.
5. X. Du, "Financial Text Analysis Using 1D-CNN: Risk Classification and Auditing Support", 2025.
6. S. Wang, R. Zhang, J. Du, R. Hao and J. Hu, "A Deep Learning Approach to Interface Color Quality Assessment in Human–Computer Interaction," arXiv preprint arXiv:2502.09914, 2025.
7. Y. Duan, L. Yang, T. Zhang, Z. Song and F. Shao, "Automated User Interface Generation via Diffusion Models: Enhancing Personalization and Efficiency," arXiv preprint arXiv:2503.20229, 2025.
8. W. Zhang et al., "Fine-tuning large language models for chemical text mining," Chemical Science, vol. 15, no. 27, pp. 10600–10611, 2024.
9. R. Zhang et al., "LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention," in Proceedings of the Twelfth International Conference on Learning Representations,pp. 106–1115, 2024.
10. B. Weng, "Navigating the landscape of large language models: A comprehensive review and analysis of paradigms and fine-tuning strategies," arXiv preprint arXiv:2404.09022, 2024.
11. L. Xia et al., "Leveraging error-assisted fine-tuning of large language models for manufacturing excellence," Robotics and Computer-Integrated Manufacturing, vol. 88, p. 102728, 2024.
12. Y. Deng, "A Reinforcement Learning Approach to Traffic Scheduling in Complex Data Center Topologies," Journal of Computer Technology and Software, vol. 4, no. 3, 2025.
13. J. Liu, Y. Zhang, Y. Sheng, Y. Lou, H. Wang and B. Yang, "Context-Aware Rule Mining Using a Dynamic Transformer-Based Framework," arXiv preprint arXiv:2503.11125, 2025.
14. 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.
15. Y. Deng, "A Hybrid Network Congestion Prediction Method Integrating Association Rules and Long Short-Term Memory for Enhanced Spatiotemporal Forecasting," Transactions on Computational and Scientific Methods, vol. 5, no. 2, 2025.
16. L. Wu, J. Gao, X. Liao, H. Zheng, J. Hu and R. Bao, "Adaptive Attention and Feature Embedding for Enhanced Entity Extraction Using an Improved BERT Model," in Proceedings of the 2024 Fourth International Conference on Communication Technology and Information Technology, pp. 702–705, December 2024.

17. Z. Yu, S. Wang, N. Jiang, W. Huang, X. Han and J. Du, "Improving Harmful Text Detection with Joint Retrieval and External Knowledge," arXiv preprint arXiv:2504.02310, 2025.
18. 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.
19. 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.
20. K. Lv et al., "Full Parameter Fine-Tuning for Large Language Models with Limited Resources," arXiv preprint arXiv:2306.09782, 2023.
21. R. He et al., "On the Effectiveness of Adapter-Based Tuning for Pretrained Language Model Adaptation," arXiv preprint arXiv:2106.03164, 2021.
22. X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," arXiv preprint arXiv:2101.00190, 2021.
23. E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in Proceedings of the International Conference on Learning Representations, Article no. 3, 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.