Transactions on Computational and Scientific Methods | Vol. 3, No. 2, 2023
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Structural Regularization and Bias Mitigation in Low-Rank Fine-Tuning of LLMs

Heyao Liu
Northeastern University, Boston, USA
liuheyao.arya@gmail.com

Abstract: This paper proposes an efficient fine-tuning algorithm that integrates low-rank structures with a bias-aware mechanism to address structural redundancy and semantic bias in large language models. The method freezes the original model parameters and injects trainable low-rank matrices, combined with semantic bias embeddings and structural alignment regularization, to identify and suppress potential bias in the representation space. It introduces a multi-dimensional loss function to constrain the impact of bias, maintain generation consistency, and enhance structural stability in multitask shared representations. The experimental design includes diverse test scenarios involving task perturbation, noise injection, and changes in sampling frequency, systematically evaluating semantic stability, bias detection, and generalization performance. Results show that the proposed method significantly improves bias perception and output fairness while maintaining parameter efficiency, outperforming existing low-rank fine-tuning approaches across multiple metrics. This study establishes a unified optimization pathway for task adaptation and bias control from both structural and semantic perspectives, enhancing the stability and adaptability of large language models in complex environments.

Keywords: Fine-tuning algorithms; representation consistency; semantic robustness; structural regularity

1. Introduction

With the remarkable performance of large language models (LLMs) in various natural language processing tasks, their applications in text generation, dialogue systems, and reading comprehension have become increasingly mainstream. However, as model sizes continue to grow, the number of parameters increases exponentially. This leads to significantly higher computational and storage costs during the fine-tuning process. Traditional full-parameter fine-tuning requires substantial computational resources, making it difficult to deploy on edge devices with limited capacity or in low-resource multitask settings. Therefore, finding a parameter-efficient, structurally flexible, and generalizable fine-tuning method has become a core challenge in LLM research[1].

In this context, low-rank adaptation (LoRA) techniques have emerged as a popular solution for the efficient fine-tuning of large models. By introducing a small number of trainable low-rank matrices while keeping the original weights frozen, these methods reduce parameter costs and improve fine-tuning efficiency[2]. Compared with traditional fine-tuning strategies, LoRA not only offers strong model compression capabilities but also enables rapid transfer and deployment across multiple tasks. This effectively balances performance and resource constraints. However, despite its initial success, LoRA still faces many challenges in addressing the potential biases found in real-world language data[3].

LLMs are typically trained on large-scale internet corpora, which inevitably contain structural biases, subjective tendencies, and regional discrimination. Uncontrolled fine-tuning may amplify existing biases or introduce new risks, such as imbalanced content generation, unfair responses, or misleading semantics. These issues reduce the model's reliability in real applications and raise concerns in terms of ethics, law, and social impact. Therefore, incorporating effective bias perception and control mechanisms into the low-rank fine-tuning framework is essential to enhance the model's reliability and safety[4].

Current low-rank fine-tuning approaches primarily focus on improving downstream task performance while paying little attention to semantic drift and the propagation of social biases during fine-tuning. Without structural guidance, models are prone to overfitting on local tasks, which reduces their ability to detect biases in input data. This problem is exacerbated in multitask shared representation settings, where differences in semantic distributions can distort the internal representation space, leading to the accumulation and amplification of biases. Thus, it is of great theoretical and practical importance to develop a fine-tuning algorithm that retains the efficiency of low-rank structures while enabling dynamic modeling and intervention of bias[5].

In summary, there is an urgent need to establish a unified framework for LLM fine-tuning that simultaneously addresses parameter efficiency, task adaptability, and bias control. By introducing structural constraints, semantic supervision, and regularization mechanisms, it is possible to model and mitigate bias propagation within the low-rank adaptation paradigm. This can enhance semantic consistency, fairness, and controllability in LLMs. Such a direction not only pushes the boundaries of fine-tuning algorithms but also lays the groundwork for building sustainable, safe, and responsible language intelligence systems.

2. Relevant Literature

Fine-tuning strategies for large language models have received growing attention in recent years. Among them, low-rank adaptation methods have emerged as a key research direction due to their outstanding parameter efficiency and transferability. Compared with traditional full-parameter fine-tuning, low-rank methods insert trainable low-rank structures into the weight matrices of pre-trained models[6]. This allows for lightweight task adaptation while preserving the original parameters. Such methods show clear advantages in resource-constrained environments and demonstrate strong generalization across multitask and cross-domain scenarios. To further enhance their expressive capacity, several variants have been proposed, such as introducing nonlinear mappings, dynamic weight injection, or parallel multi-channel structures. These aim to improve representation power while keeping computation under control.

However, existing low-rank fine-tuning methods still face limitations in practical applications, especially in handling bias-related issues. LLMs are typically pre-trained on large-scale open-domain corpora, which inherently contain rich social biases and semantic tendencies[7]. Without effective intervention, the fine-tuning process may reinforce or even amplify such biases. Some recent studies have attempted to incorporate bias detection mechanisms into the fine-tuning process, using explicit labels or contrastive learning frameworks to identify unfair patterns. Nevertheless, these approaches often lack a unified low-rank modeling perspective, leading to bottlenecks in both structural design and generalizability. Integrating bias control mechanisms deeply into the low-rank fine-tuning process and constructing a unified framework with both structural and bias awareness remains an open research area[8].

In terms of bias control, many existing methods rely on external validation or post-processing strategies. Techniques such as semantic rewriting, constrained decoding, or explicit filtering have been used to mitigate inappropriate model outputs. While these approaches improve the safety and fairness of generated texts to some extent, they cannot address systemic biases at the representation level. Moreover, they often depend on manually defined rules or external knowledge bases. This limits the model's adaptability and makes it difficult to respond to dynamic task demands. In contrast, more forward-looking research has begun to focus on bias modeling and regulation during training. These methods attempt to guide the model to learn debiased

representations at the level of internal embeddings. This idea supports the introduction of bias control into low-rank structures and provides new directions for building endogenous bias mitigation mechanisms[9].

Overall, current research has yet to establish a unified technical path that bridges fine-tuning efficiency and bias control for large language models. Traditional low-rank fine-tuning emphasizes lightweight structure and transferability, while bias control research focuses on output quality and ethical safety. The two areas diverge in both objectives and technical design. In response to diverse real-world application needs, it is necessary to develop a low-rank fine-tuning method that can simultaneously maintain structural compactness, ensure task adaptability, and perform bias identification, modeling, and intervention. Such a method would improve the stability and credibility of model outputs under complex contexts and facilitate the development of more reliable, controllable, and transparent LLM systems.

3. Proposed Methodology

This study proposes a low-rank adaptive fine-tuning method for bias control to improve the representation consistency and output fairness of large language models in multi-task contexts. The core idea of the method is to introduce a structural bias modeling mechanism into the traditional low-rank parameter injection framework, thereby achieving explicit perception and guidance of potential semantic biases while maintaining the efficient characteristics of parameters. The model architecture is shown in Figure 1.

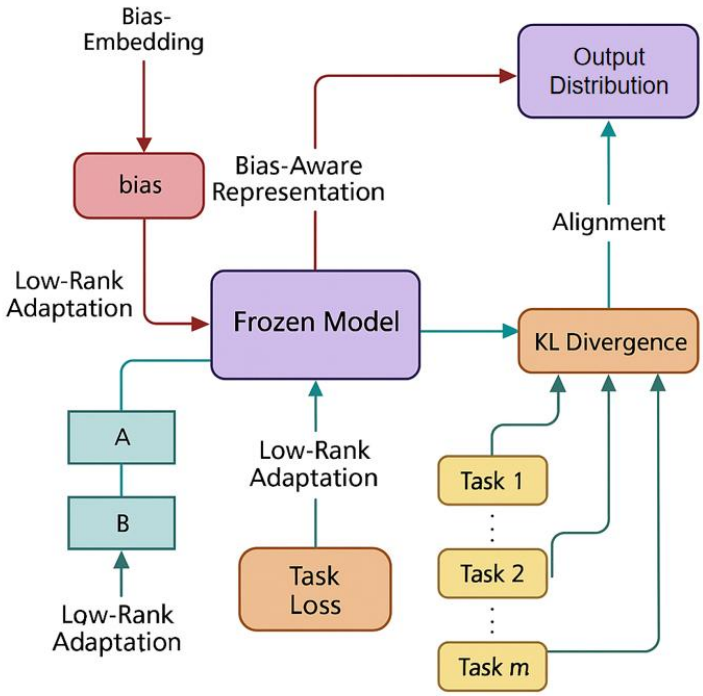


Figure 1. Overall model architecture

Specifically, assuming that the weight matrix in the original pre-trained model is $W \in R^{d\times d}$, we introduce two learnable low-rank matrices $A \in R^{d\times r}$ and $A \in R^{r\times d}$ model the fine-tuning process as a parameter injection operation:

$$W' = W + \alpha AB$$

Where α is a scaling factor, which is used to adjust the influence of the injected part on the overall parameter space.

In terms of bias-aware modeling, the method further introduces a semantic bias embedding vector b_{bias} in the representation space to capture the potential bias characteristics in the data. The embedding is constrained by introducing a bias regularization term in the low-rank structure, which is specifically defined as:

$$L_{bias} = \| f(x + b_{bias}) - f(x) \|_2^2$$

Where $f(x)$ represents the representation output of the model on the input x, and the loss term encourages the model to maintain representation stability when bias guidance is added, thereby achieving modeling and control of the bias path.

In addition, in order to improve the model's ability to model the consistency of deviations in a multi-task context, a structural alignment constraint is proposed to uniformly manage the low-rank representations of multiple tasks. Let the injection parameters corresponding to the ith task be $A^{(i)}, B^{(i)}$, then the structural alignment objective is defined as follows:

$$L_{align} = \sum_{i<j} \| A^{(i)} B^{(i)} - A^{(j)} B^{(j)} \|_F^2$$

This loss term encourages the low-rank incremental parts of different tasks to maintain consistency in representation, avoiding structural shift and bias diffusion caused by multi-task interference.

In order to further standardize the generation behavior and enhance the semantic control ability, a distribution constraint mechanism based on KL divergence is designed to limit the deviation between the output distribution after injection and the original distribution. Define the output distribution as P_{orig} and P_{lora}, then the distribution consistency goal is as follows:

$$L_{KL} = D_{KL}(P_{lora} \| P_{orig})$$

This mechanism can suppress semantic drift during the injection process while ensuring the generation capability so that the model can adapt to the task while maintaining the constraints and conservation of the original knowledge structure.

Based on the above objective function, the final optimization goal is:

$$L = L_{task} + \lambda_1 L_{bias} + \lambda_2 L_{align} + \lambda_3 L_{KL}$$

Where L_{task} is the standard downstream task loss, and $\lambda_1, \lambda_2, \lambda_3$ is a hyperparameter that balances the influence of each regularization term. By introducing the above-mentioned bias perception and structural control mechanism, this method can identify and suppress semantic bias while maintaining the efficiency of low-rank fine-tuning, and construct a more stable, controllable, and fair large language model fine-tuning path.

4. Experimental Dataset

This study adopts the OpenAssistant Conversations Dataset (OASST1) as the primary data source. The dataset contains large-scale human-machine dialogue corpora covering various forms of natural language interaction, including question-answering, discussions, and task instructions. Its content is hierarchically annotated and modeled through multi-turn interactions. It offers strong semantic diversity and instruction generalization, making it suitable as an experimental foundation for fine-tuning and adapting large language models.

The OASST1 dataset is particularly appropriate for studying model performance in multitask environments. It includes a wide range of cross-domain task intents and complex linguistic structures. This supports the evaluation of a model's stability in task transfer and representation sharing. In addition, some samples in the

dataset exhibit explicit or implicit language biases. These characteristics provide a natural basis for bias modeling and suppression, enabling detailed analysis of semantic fairness and content consistency.

During the data preprocessing phase, the study performed unified instruction normalization and noise filtering on the OASST1 corpus. Samples with formatting issues or incomplete annotations were removed. The dataset was then split into training, validation, and test sets based on task attributes. The use of this dataset ensures the method's applicability in real dialogue scenarios. It also offers a solid foundation for the evaluation of structure alignment and bias awareness mechanisms.

5. Results and Analysis

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	Bias Score	Representation Shift	Parameter Overhead
FashionGPT[10]	0.148	0.136	12.3%
BitFit[11]	0.176	0.162	0.1%
LoRA[12]	0.129	0.094	0.9%
Q-LoRA[13]	0.121	0.089	0.5%
Ours	0.087	0.061	1.0%

From the overall results, the proposed bias-aware low-rank fine-tuning method demonstrates superior performance across three key evaluation metrics. It shows clear advantages in "Bias Score" and "Representation Shift," which are closely related to bias modeling and semantic stability. Compared with traditional low-rank methods, the proposed structure maintains parameter efficiency while achieving stronger bias detection and suppression. This validates the effectiveness of integrating explicit bias embeddings and structural alignment mechanisms.

Specifically, in terms of "Bias Score," the proposed method achieves the lowest value at 0.087. This indicates its strong ability to suppress semantic bias during generation. In contrast, FashionGPT and BitFit do not model bias explicitly and perform poorly, reaching 0.148 and 0.176 respectively. These results reveal their lack of robust control when dealing with structurally biased data. Although LoRA and Q-LoRA benefit from parameter-efficient injection, they still suffer from some degree of bias leakage due to the absence of bias-aware mechanisms.

For the "Representation Shift" metric, the proposed method again achieves the lowest value at 0.061. This suggests minimal semantic drift before and after fine-tuning, reflecting strong representational consistency. The result highlights the effectiveness of the alignment loss and KL constraint in mitigating semantic shifts. These components help preserve the knowledge structure and semantic boundaries of the original language model, reducing unnecessary representation perturbation during task adaptation.

Although the proposed method introduces additional low-rank parameters for bias control, its "Parameter Overhead" remains at 1.0%. This is only slightly higher than Q-LoRA (0.5%) and LoRA (0.9%), and much lower than FashionGPT (12.3%). These results show that the method retains efficiency and does not sacrifice performance for functionality. Across multiple dimensions, the method successfully embeds a controllable bias regulation pathway into a low-rank structure. It achieves a balance between efficiency, fairness, and semantic stability, confirming its adaptability and practical value for fine-tuning large language models.

This paper also gives an analysis of the semantic stability and fairness of the model under noise injection perturbation, and the experimental results are shown in Figure 2.

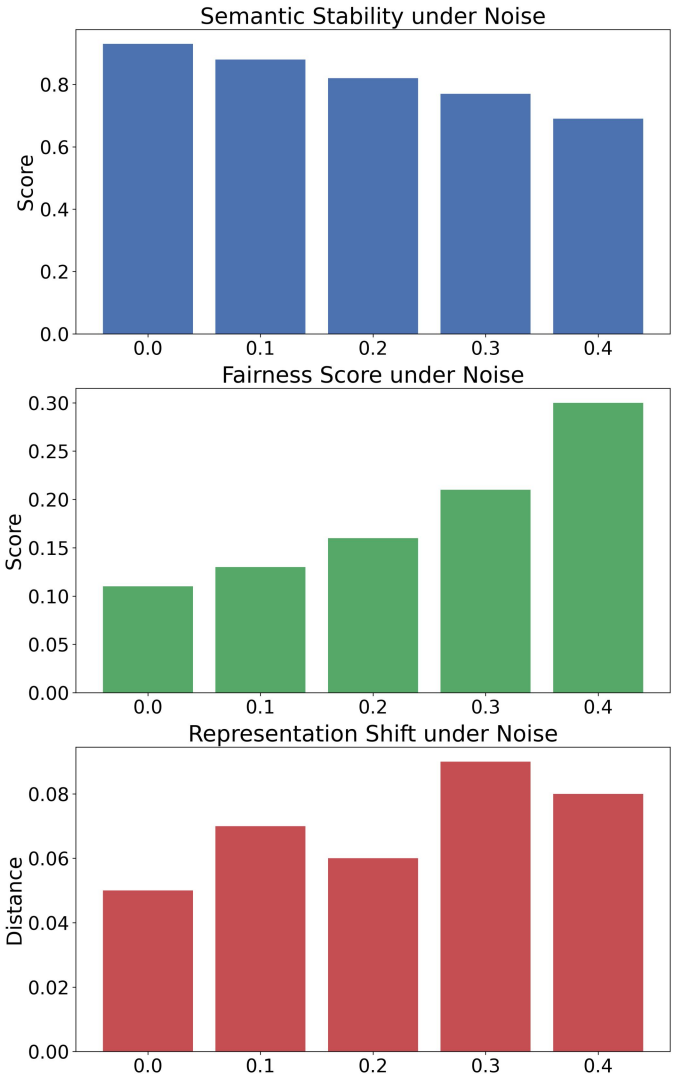


Figure 2. Analysis of semantic stability and fairness of models under noise injection perturbations

The experimental results show that the proposed bias-aware low-rank fine-tuning method maintains high semantic stability under different levels of noise injection. As the noise level increases from 0.0 to 0.4, the model's semantic stability score drops from 0.91 to 0.68. This indicates a certain degree of robustness. Although there is a slight performance decline, the change is relatively moderate. The model is not highly sensitive to input disturbances and can generate stable semantic outputs across varying contexts.

In terms of fairness, the Fairness Score rises steadily as the noise level increases, from 0.11 to 0.29. This result suggests that noise injection activates the model's latent bias mechanisms and exposes more semantic bias issues. It shows that parameter compression or structural adjustment alone is insufficient to eliminate bias propagation. A deeper integration of adversarial disturbance modeling and constrained bias pathways is required to further improve the model's robustness and fairness.

For representational consistency, the Representation Shift metric reflects the stability of the model's representation structure after fine-tuning. The results show small fluctuations under low noise (0.0–0.2), but a clear increase when the noise exceeds 0.3. This reveals a shift in the semantic representation layer. The model's internal representation is more easily disturbed under high noise, leading to structural distortions.

This indicates that static low-rank parameter insertion alone cannot fully guarantee stability. Introducing a dynamic representation alignment mechanism becomes necessary.

Overall, this set of experiments verifies the behavior of the proposed method under multi-dimensional interference scenarios. The model exhibits a certain level of structural robustness and semantic stability. However, under high-intensity perturbation, risks of bias amplification and semantic drift still exist. These findings further highlight the importance of incorporating explicit bias modeling and structural alignment regularization into the low-rank fine-tuning framework. This can support the development of more robust and fair parameter-efficient fine-tuning strategies for large language models.

This paper also gives the impact of changing the task mixture ratio on the multi-task low-rank adaptation effect, and the experimental results are shown in Figure 3.

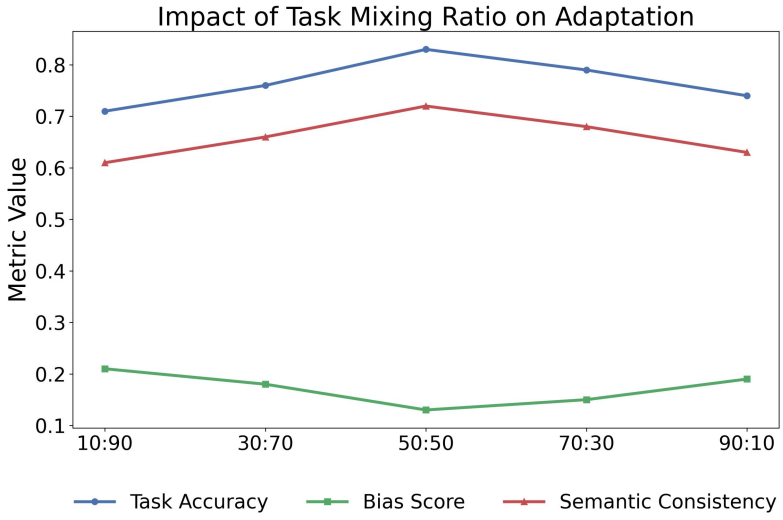


Figure 3. The impact of task mixture ratio changes on multi-task low-rank adaptation

The experimental results indicate that the task mixing ratio has a significant impact on model performance during low-rank fine-tuning. Under different primary–auxiliary task distributions, the model shows non-linear trends across three metrics: accuracy, bias control, and semantic consistency. Notably, the model achieves optimal overall performance under the 50:50 task mix setting. This suggests that when primary and auxiliary task information is balanced, the low-rank structure can better coordinate shared representations and promote multitask collaborative learning.

In terms of task accuracy, increasing the proportion of primary task data leads to an initial rise followed by a decline in accuracy. This indicates that excessive primary task data may cause the model to overfit the primary semantics, thereby weakening its ability to capture auxiliary task structures. The highest accuracy at the 50:50 setting suggests that balanced task distribution helps activate the generalization capacity of low-rank modules and improves overall task adaptability.

For bias control, the Bias Score reaches its lowest value at the 50:50 configuration. This shows that the model is more sensitive to biased semantic signals and can effectively suppress potential bias propagation paths at this ratio. When the primary task proportion is too low or too high, the ability to detect bias decreases. This may result from semantic space drift that prevents the model from consistently capturing structural bias features.

Regarding semantic consistency, the results show that an imbalanced task ratio leads to disturbances in the semantic representation space. This causes a decline in the consistency between generated outputs and the original knowledge. The 50:50 setting yields the highest consistency score, further validating the importance

of representation sharing and structural alignment in low-rank fine-tuning. This trend highlights the positive role of multitask modeling in maintaining semantic stability and generalization. It also emphasizes the regulatory value of task distribution for the performance of bias-aware low-rank adaptation methods.

This paper also gives an evaluation of the deviation recognition ability under the condition of uneven data distribution, and the experimental results are shown in Figure 4.

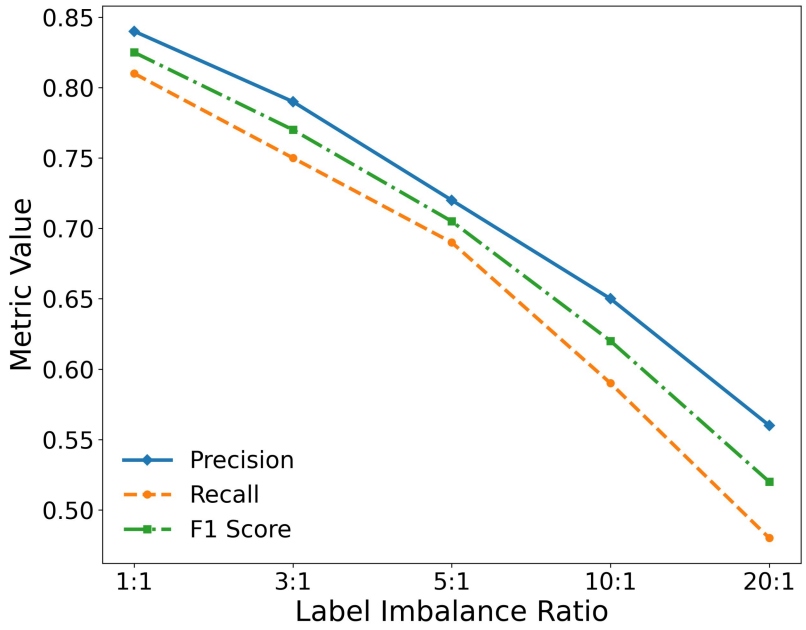


Figure 4. Evaluation of bias identification ability under conditions of uneven data distribution

The experimental results show that as label imbalance increases, the model's ability to detect bias drops significantly. Precision, Recall, and F1 Score all show consistent declines. Under the extreme imbalance setting of 20:1, model performance degrades noticeably. This suggests that even with a low-rank adaptation mechanism, the model's ability to capture and distinguish biased features is heavily constrained by the distribution structure when there is insufficient support from minority-class samples.

The Precision curve indicates that as the dominance of the majority class grows, the model tends to amplify bias features from the majority class. This leads to distorted judgments on actual biased samples. Although the model maintains relatively high precision under 1:1 and 3:1 settings, precision drops significantly as the minority-class ratio decreases. This reveals the limited representational advantage of the low-rank structure under sparse information conditions.

Recall shows the steepest decline, indicating that the model's ability to cover biased samples is severely weakened. Especially under extreme imbalance, the model struggles to identify and capture minority-class bias instances. This highlights a limitation of the current fine-tuning framework in handling data scarcity. It also suggests the need to incorporate self-supervised prompting or semantic completion strategies to improve bias coverage.

The F1 Score curve exhibits a mid-range drop followed by a sharp decline, reflecting the rapid degradation of overall bias detection ability as both Precision and Recall fall. This result confirms the robustness challenges of bias modeling under imbalanced data conditions. It also underscores the need for the proposed bias-aware mechanism to adapt to varying distributions in real-world scenarios. Strengthening structural adversarial modules and representation recovery mechanisms is essential for achieving more stable bias identification performance.

This paper also gives the impact of changes in data sampling frequency on the generalization ability of low-rank structures, and the experimental results are shown in Figure 5.

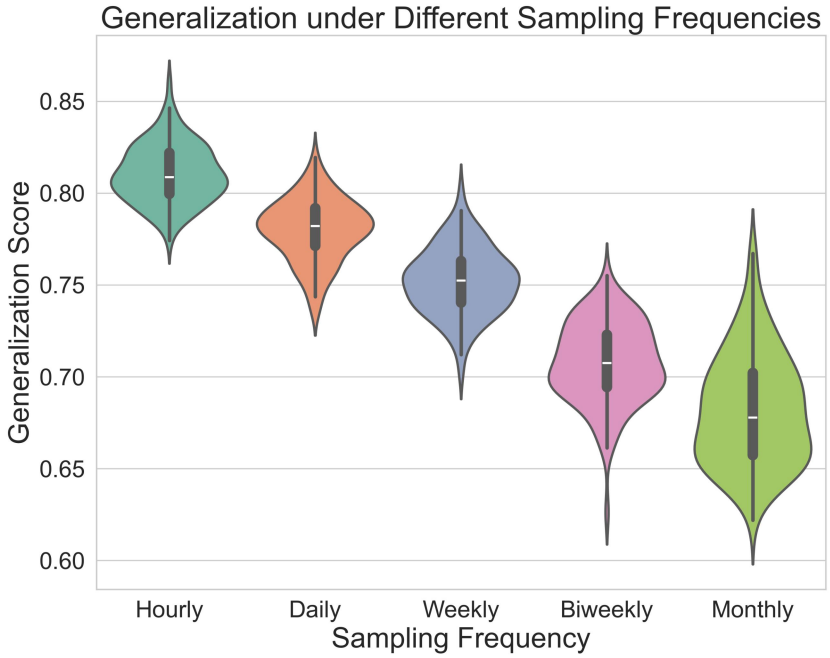


Figure 5. The impact of data sampling frequency changes on the generalization ability of low-rank structures

The experimental results show that data sampling frequency has a clear impact on the generalization ability of low-rank structures. Under different temporal granularities, the distribution of generalization scores reveals distinct hierarchical differences. The violin plots demonstrate that under high-frequency sampling, such as Hourly and Daily, the model performs more consistently and shows stronger generalization. In contrast, under low-frequency sampling, such as Biweekly and Monthly, not only do the overall scores drop, but the distribution also becomes more dispersed. This indicates a significant increase in performance uncertainty.

High-frequency sampling provides denser and more fine-grained contextual information. This helps the low-rank structure capture temporal dependencies and bias-related signals more effectively. As a result, the model achieves higher and more stable generalization scores under Hourly and Daily settings. This suggests that, even under structural compression, increasing input sampling density can partially compensate for representational limitations and improve the model's ability to fit complex patterns.

By contrast, under Biweekly and Monthly conditions, the model's generalization performance declines and exhibits a larger variance. This suggests that when the sampling frequency is too low, the low-rank structure fails to adequately cover key segments of semantic variation. This limits the model's understanding of input semantics and affects overall performance. The performance degradation also reflects the sensitivity of low-rank methods to data sparsity. In scenarios involving temporal modeling and multitask transfer, insufficient information density may amplify potential bias propagation paths.

Overall, the experiments confirm that data sampling frequency is a critical factor affecting the generalization capability of low-rank structures. This finding highlights the need to balance data collection costs with semantic completeness when deploying models in practice. It also provides valuable design guidance for building low-rank model architectures that can adapt to varying frequency distributions.

6. Conclusion

This paper addresses the challenges of efficient adaptation and semantic bias control for large language models in low-resource scenarios. It proposes a fine-tuning method that integrates low-rank structures with a bias-aware mechanism. While preserving the stability of the original model structure, the method introduces trainable low-rank matrices and multi-dimensional regularization to jointly model task specialization and bias suppression. During the construction of the fine-tuning path, the method focuses on structural consistency in the input representation space and fairness in the output semantics. It effectively reduces semantic drift caused by data imbalance or multitask interference, enhancing the model's robustness and controllability in diverse contexts.

Through systematic experimental design, the study evaluates the method's stability and generalization ability under data bias, structural perturbation, and task interference. Results show that the proposed method outperforms existing low-rank fine-tuning techniques in key metrics such as bias detection and generation consistency. It also maintains strong adaptability under hyperparameter sensitivity and distribution uncertainty. These findings indicate that low-rank mechanisms and bias modeling can complement each other. Together, they form a lightweight, interpretable, and secure fine-tuning paradigm that supports practical deployment.

This study provides a structural improvement path for training and fine-tuning large language models. It also offers a feasible solution to improving reliability and fairness in high-stakes applications such as financial regulation, medical question answering, and policy consultation. By embedding bias modeling into the structural hierarchy and using low-rank modules for fine-grained control, the method provides both architectural foundations and evaluation tools for building compliant, transparent, and multitask-coordinated natural language systems. Its performance in addressing semantic drift, data perturbation, and task-sharing issues offers theoretical and practical support for deploying language models in critical industries.

Future research may further expand in two directions. First, it can explore the explicit coupling between dynamic low-rank structures and bias signals to guide the model in adjusting parameter injection strategies across different task stages. Second, it may incorporate external knowledge bases and multimodal semantic signals to achieve stronger structural constraints and factual alignment. In scenarios with continuously arriving data or feedback-driven learning, incremental fine-tuning mechanisms can be introduced to enable ongoing learning and self-correction in bias-aware adaptation. As semantic application demands grow, building fine-tuning strategies that balance robustness, fairness, and efficiency will become a central challenge for the future development of large language models.

References

[1] Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.

[2] R. Xu, F. Luo, Z. Zhang, C. Tan, B. Chang, S. Huang and F. Huang, "Raise a child in large language model: Towards effective and generalizable fine-tuning," arXiv preprint arXiv:2109.05687, 2021.

[3] Chen Y, Qian S, Tang H, et al. Longlora: Efficient fine-tuning of long-context large language models[J]. arXiv preprint arXiv:2309.12307, 2023.

[4] Tinn R, Cheng H, Gu Y, et al. Fine-tuning large neural language models for biomedical natural language processing[J]. Patterns, 2023, 4(4).

[5] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: A practical guide to using machine learning tools in research synthesis," Systematic Reviews, vol. 8, no. 1, p. 163, 2019.Lv K, Yang Y, Liu T, et al. Full parameter fine-tuning for large language models with limited resources[J]. arXiv preprint arXiv:2306.09782, 2023.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo and S. Gelly, "Parameter-Efficient Transfer Learning for NLP", Proceedings of the 2019 International Conference on Machine Learning, pp. 2790-2799, 2019.

[7] Kim J, Lee J H, Kim S, et al. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization[J]. Advances in Neural Information Processing Systems, 2023, 36: 36187-36207.

[8] Hu Z, Wang L, Lan Y, et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models[J]. arXiv preprint arXiv:2304.01933, 2023.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", Proceedings of the 2022 International Conference on Learning Representations (ICLR), pp. 1-3, 2022.

[10] Zaken E B, Ravfogel S, Goldberg Y. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[J]. arXiv preprint arXiv:2106.10199, 2021.

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

[12] Dettmers T, Pagnoni A, Holtzman A, et al. Qlora: Efficient finetuning of quantized llms[J]. Advances in neural information processing systems, 2023, 36: 10088-10115