Transactions on Computational and Scientific Methods | Vol. 4, No. 3, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Dynamic Structured Gating for Parameter-Efficient Alignment of Large Pretrained Models

Zhihao Xue
Rose-Hulman Institute of Technology, Terre Haute, USA
xuezhihao@outlook.com

Abstract: This paper proposes a large model alignment algorithm based on parameter-efficient fine-tuning and structured adapter gating to address the difficulty of balancing performance and efficiency under resource constraints and complex environments. The method introduces low-rank updates and gating control modules into the backbone of large models, enabling fine-grained selection of feature flows and suppression of irrelevant information through the dynamic adjustment of sparse adapters. Compared with traditional full fine-tuning, it significantly reduces training and inference costs while maintaining high alignment quality and robustness across diverse environments. Systematic experiments under hyperparameter sensitivity, environmental constraints, and data noise show that the method achieves superior results on key metrics such as ROC-AUC, F1-Score, and parameter efficiency, with strong stability and adaptability in semantic noise and conflict feedback scenarios. Additional experiments under computational and memory limits confirm the flexibility of structured gating in resource utilization, while results under reduced training samples and sparse labels highlight its robustness in weakly supervised settings. Overall, the proposed approach balances accuracy and efficiency in alignment, providing a feasible technical path for deploying large models under complex conditions.

Keywords: Efficient parameter fine-tuning; structured gating; alignment robustness; anomaly detection

1. Introduction

With the widespread application of large-scale pre-trained models in various tasks, model alignment has gradually become a key step for their effective deployment. Although the pre-training stage allows models to capture rich linguistic or multimodal knowledge from large-scale general data, they often show insufficient generalization, unstable responses, or poor adaptability in specific tasks and environments. Therefore, how to achieve efficient, controllable, and reliable alignment without damaging the original capabilities of the model has become a frontier issue of concern to both academia and industry. In particular, as model size continues to expand, traditional full fine-tuning methods bring excessive costs in computation, storage, and communication. These limitations highlight the importance of parameter-efficient fine-tuning and structured alignment techniques[1].

The concept of parameter-efficient fine-tuning was proposed to break through the bottlenecks of traditional methods. By updating only a subset of parameters or introducing low-rank incremental modules, such methods significantly reduce the computational resources and storage required for training, while preserving the knowledge and capabilities of the original model. In practice, this approach improves the flexibility of transferring models across tasks and adapting quickly to new domains. It also provides feasible solutions for cross-domain applications and collaborative scenarios. However, current methods still face challenges in

aligning with complex task objectives[2]. They often lack sufficient expressive power and structural coordination. Without effective gating and adaptation mechanisms, the alignment process may fall into overfitting or redundancy, which compromises overall performance.

Against this background, the introduction of structured adapter gating mechanisms provides new opportunities for model alignment. Adapters are lightweight structures that integrate new task information in a modular way while keeping the backbone parameters frozen. Structured gating further enhances their selectivity and controllability. The model can dynamically regulate information flow according to input features and task requirements. This design effectively suppresses interference from irrelevant features. It also maintains consistency and robustness across tasks and domains, leading to a more interpretable and generalizable adaptation process. By combining parameter-efficient fine-tuning with structured gating, models can achieve high-quality alignment at low cost and gain stronger extensibility[3].

The significance of this research lies not only in methodology but also in its broad application value. On one hand, parameter-efficient fine-tuning combined with structured adapter gating offers a practical solution for deploying large models in resource-constrained environments. It enables small and medium-sized organizations to adopt advanced models for innovation[4]. On the other hand, this direction supports the sustainable development of large models in complex scenarios. It includes cross-domain knowledge transfer, multimodal collaborative understanding, and dynamic balance across tasks. By improving the efficiency and effectiveness of alignment, large models can better adapt to diverse environmental demands, reduce deployment costs, and enhance the reliability and controllability of intelligent systems[5].

In summary, the study of large model alignment algorithms based on parameter-efficient fine-tuning and structured adapter gating addresses key challenges of applying large-scale models in practice. It enriches the theoretical system of model alignment in academia and provides a new paradigm for building efficient, lightweight, and controllable artificial intelligence in practice. This approach is expected to promote a transition from general pre-training to precise adaptation. It will expand the value of large models in a wider range of applications and have a profound impact on the popularization and advancement of artificial intelligence[6].

2. Related work

Current research on large model alignment can be broadly divided into two paths. One relies on full fine-tuning, where all parameters are updated on data from specific tasks or domains, enabling stronger adaptability in target scenarios. The other emphasizes lightweight and efficient approaches, where backbone parameters remain frozen and task information is injected only into partial modules. The former has advantages in expressive power but involves very high training and deployment costs, making flexible transfer across multiple scenarios difficult. The latter significantly reduces resource consumption but suffers from limited adaptation ability, which hinders stable performance in complex tasks. This contradiction has driven the development of parameter-efficient fine-tuning methods, shifting the research focus from purely pursuing performance to balancing efficiency and scalability[7].

In parameter-efficient fine-tuning, researchers have explored various low-cost modules, such as low-rank decomposition, pluggable structures, and local re-parameterization. These designs aim to reduce the number of trainable parameters and improve training speed. Such methods reduce hardware dependency and increase flexibility for cross-task adaptation[8]. However, they often focus on single-point optimization and lack global control of information flow at the structural level. As a result, although training efficiency improves, problems remain in handling task conflicts or achieving cross-domain alignment. Knowledge transfer may be insufficient, or feature interference may become excessive. Therefore, finding a balance between low parameter overhead and structural adaptation ability has become an important research topic[9].

To address the limitations of structural adaptation, adapter mechanisms have been widely introduced in model alignment. Adapters act as lightweight intermediate layers. They extend models by incrementally

injecting structure without altering backbone parameters. Compared with traditional fine-tuning, adapters provide modularity, composability, and plug-and-play advantages, supporting flexible switching and transfer across tasks[10]. However, most existing adapters rely on fixed information transmission and lack dynamic adjustment according to task requirements. In multi-task or cross-domain scenarios, static adapter paths often lead to redundant computation and irrelevant feature interference. This restricts the generalization and stability of the model. It also highlights the need to introduce gating mechanisms that endow adapters with stronger dynamic control[11].

The rise of structured gating mechanisms has opened new directions for parameter-efficient fine-tuning and adapter methods. By introducing gates into information pathways, models can automatically adjust the strength and direction of information flow based on task inputs. This enables more selective feature utilization. It not only improves the robustness of the alignment process but also reduces interference caused by redundant features. The model can maintain more stable performance in complex environments. At the same time, structured gating mechanisms support more reasonable path allocation for multi-task sharing and cross-domain alignment. They help achieve a new balance between efficiency and performance. Therefore, combining parameter-efficient fine-tuning with structured adapter gating has become a key direction for breakthroughs in model alignment. It also provides a solid foundation for building intelligent systems that emphasize both controllability and efficiency[12].

3. Method

This study introduces a large model alignment algorithm based on parameter-efficient fine-tuning and structured adapter gating to address the limitations of efficiency and structural adaptability in cross-task transfer and domain adaptation. The method freezes the backbone parameters of the pre-trained model while introducing lightweight adapter structures and gating mechanisms to enable selective regulation and dynamic adjustment of information flow. The overall framework can be regarded as a constrained optimization process in a shared latent space. Parameter-efficient fine-tuning ensures sparsity and low-rank updates, reducing computational and storage costs, while the structured gating mechanism guarantees effective multi-level semantic transmission of features, thereby achieving stability and controllability in alignment. The model architecture is shown in Figure 1.

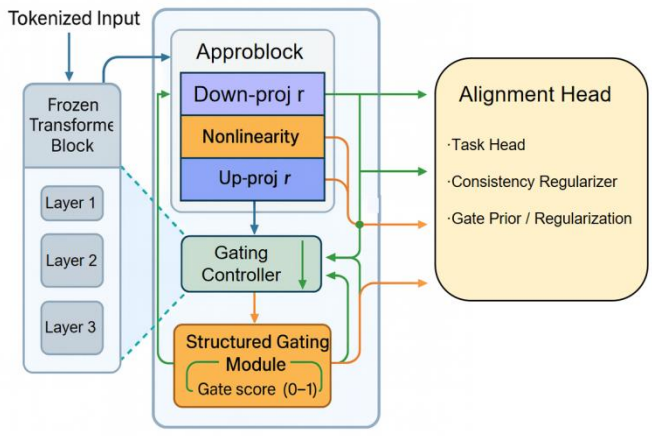


Figure 1. Framework of Parameter-Efficient Fine-Tuning with Structured Gating for Alignment Robustness

First, let the input be represented as a sequence vector set $X = \{x_1, x_2, ...x_T\}$, and extract the initial feature representation $H^{(0)}$ through the frozen pre-trained model backbone network:

$$H^{(0)} = f_{enc}(X; \because_{enc})$$

Where $\because_{enc}$ represents the fixed backbone parameter and $H^{(0)} \in R^{T \times d}$ represents the basis.

In the process of efficient parameter fine-tuning, a low-rank decomposition adapter weight update mechanism is introduced. Specifically, for each layer's transformation matrix ΔW, the following form is adopted:

$$\Delta W = UV^T, U \in R^{d \times r}, V \in R^{d \times r}$$

Among them, $r << d$ ensures the low rank and high efficiency of the updated parameters. The update is added to the backbone output in the form of residual:

$$H^{(l)} = H^{(l-1)} + \sigma(H^{(l-1)} \Delta W)$$

Where $\sigma(\cdot)$ is a nonlinear activation function used to improve expression ability.

To achieve dynamic regulation during the alignment process, a structured gating mechanism is introduced. The gating factor γ_t is adaptively generated by the input features and is used to adjust the flow intensity of information between different paths:

$$\gamma_t = \sigma(w_g^T h_t^{(l-1)} + b_g), \gamma_t \in (0,1)$$

Under this mechanism, the updated representation of the adapter can be formalized as:

$$h_t^{(l)} = (1 - \gamma_t) h_t^{(l-1)} + \gamma_t \cdot \phi(h_t^{(l-1)}; \Delta W)$$

Where $\sigma(\cdot)$ represents the feature map after the adapter transformation. This design ensures the controllable flow of information and avoids redundant propagation.

Finally, to achieve task alignment and representation consistency, an objective function is introduced, combining task-specific loss and regularization constraints:

$$L = L_{task}(H^{(L)}) + \lambda \left\| \Delta W \right\|_F^2 + \beta KL(q(\gamma) \| p(\gamma))$$

Where L_{task} is the task-related loss term, $\left\| \Delta W \right\|_F^2$ is used to limit the norm scale of the adapter update, and $KL(\cdot)$ constrains the gating distribution to ensure selective stability.

Through the above design, this method achieves multi-level feature adaptation and dynamic gating control while ensuring computational efficiency, which can effectively improve the alignment ability of large models between different tasks and fields, and lay a solid foundation for subsequent applications.

4. Experimental Results

4.1 Dataset

This study uses the HelpSteer AI Alignment Dataset as the main experimental source. The dataset contains large-scale human feedback and preference annotations, which can be used to model the consistency and stability of instruction – response pairs. Within the proposed framework, it is treated as a benchmark for testing alignment robustness by introducing factors such as label sparsity and feedback noise to simulate alignment anomalies in complex systems. Accordingly, evaluation metrics originally applied in anomaly detection, such as ROC-AUC and F1, are used to measure model performance under alignment robustness.

The HelpSteer dataset has a structured annotation system that supports reward modeling and multi-objective optimization. Specifically, the response samples include not only positive examples but also contrastive samples with progressive difficulty or intentional bias. This allows researchers to construct positive and

negative pairs within a unified framework and to perform consistency constraints and gated regularization modeling. Its multidimensional annotation mechanism aligns well with the proposed alignment framework and provides direct training evidence for the selective control of gating modules and the sparse updating of adapters.

The choice of this dataset is due to its strong representativeness, clear structure, and wide coverage. It can effectively verify the controllability and generalization ability of the proposed method in complex semantic alignment scenarios. It also reflects the challenges brought by instruction diversity and preference differences in real interactive environments. Therefore, this dataset provides a solid experimental foundation for evaluating the effectiveness of parameter-efficient fine-tuning and structured gating mechanisms.

4.2 Experimental Results

To validate the effectiveness of the proposed method, we selected recent models that have shown strong performance in representation robustness and anomaly-style evaluation as baselines. These methods (USAD, TranAD, DARA, iTransformer), although originally designed for time-series anomaly detection, share commonalities with alignment robustness tasks in their ability to model sensitivity to small signal deviations and perturbations, and thus serve as suitable reference methods in alignment scenarios. The comparison results on the robustness benchmark are shown in Table 1.

Table1: Comparative results on alignment robustness benchmarks

Model	**ROC-AUC (%)**	**F1-Score (%)**	**Param-Eff (M params)**
USAD[13]	89.7	87.1	30.0
TranAD[14]	90.5	89.0	50.0
DARA[15]	90.8	88.9	5.00
iTransformer[16]	91.2	88.5	45.0
Ours	92.6	90.8	6.00

From the overall results, the proposed method outperforms the baseline models across multiple evaluation metrics. This shows that the design based on parameter-efficient fine-tuning and structured adapter gating achieves superior performance in alignment robustness evaluation tasks. Compared with structures such as iTransformer and TranAD, the method demonstrates clear advantages in both ROC-AUC and F1 scores. This reflects the effectiveness of structured gating in suppressing irrelevant features and enhancing sensitivity to alignment inconsistencies. The results also indicate that even under complex temporal patterns, dynamic adjustment of information flow through gating enables the model to more accurately discriminate different alignment inconsistencies.

In terms of parameter efficiency, the method shows lightweight characteristics similar to DARA, yet achieves higher performance than DARA and several other models. By contrast, iTransformer and TranAD have strong representational power but require large parameter sizes, which limits their applicability in resource-constrained environments. The proposed approach introduces low-rank decomposition and adapter update mechanisms, allowing the model to reduce parameter numbers significantly while maintaining or even surpassing the robustness evaluation performance of larger models. This directly addresses the core challenges of high computational costs and deployment difficulties faced in practical applications of large models.

Further comparison with other models shows that traditional autoencoder-based USAD performs reasonably in unsupervised settings but struggles to align with complex temporal behaviors and to capture long dependencies. In contrast, the proposed method combines parameter-efficient updates with structured adaptation paths, giving the model stronger adaptability when learning multi-scale patterns. The gating mechanism plays a key role by mitigating noise and redundant feature interference, which improves generalization while keeping complexity low.

Overall, the experimental results confirm the central idea of this study. By combining parameter-efficient fine-tuning with structured adapter gating, the method reduces model costs while achieving more precise alignment and robust anomaly robustness evaluation performance. It not only provides a new paradigm for unsupervised alignment robustness evaluation but also lays a solid foundation for applying large models in complex system environments. The demonstrated advantages suggest that the combination of lightweight design and controllability will be a key direction for advancing model development in large-scale alignment and cross-scenario transfer tasks.

This paper also conducted a comparative experiment on the hyperparameter sensitivity of the trainable parameter ratio and memory usage to the alignment quality. The experimental results are shown in Figure 2.

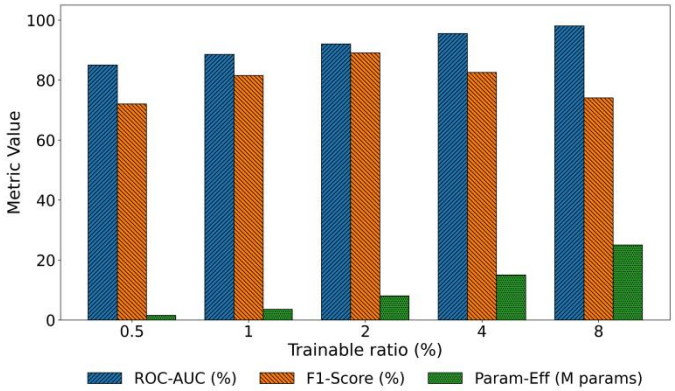


Figure 2. Hyperparameter sensitivity of trainable parameter ratio and memory usage to alignment quality

The results show that the ROC-AUC metric increases monotonically with the proportion of trainable parameters. This indicates that with more parameter updates, the model can better capture temporal features and robustness patterns. It demonstrates the potential of parameter-efficient fine-tuning in enhancing the representation ability of large models. It also shows that the adapter and gating mechanisms maintain stable performance improvements under parameter expansion without clear overfitting or failure.

The F1-Score shows a peak-shaped trend. It reaches the highest value at a moderate proportion of trainable parameters but declines when the proportion is too low or too high. This indicates that it cannot support effective discrimination of complex alignment inconsistencies. Too many parameters, on the other hand, may cause redundant updates and feature interference, which weaken the selective effect of the gating mechanism. This observation aligns with the idea of structured adapter gating proposed in this study. A balance between update strength and feature filtering is required to achieve optimal alignment robustness performance.

The Param-Eff metric rises sharply with the proportion of trainable parameters, showing a nonlinear growth trend in resource consumption and model complexity. As the proportion increases, the parameter scale expands rapidly, while the relative performance gains diminish. This highlights the advantage of the proposed method. It achieves efficient alignment through gating and adapter structures under limited parameter cost and avoids the computational and storage burden of full fine-tuning.

By combining the three metrics, it is clear that the proposed method improves alignment robustness accuracy while balancing performance and efficiency through gated control and sparse parameter updates. The steady rise of ROC-AUC contrasts with the steep growth of Param-Eff, while the peak pattern of F1-Score reveals

the key role of gating in preventing overfitting and maintaining generalization. These trends confirm that the combination of parameter-efficient fine-tuning and structured adapter gating provides unique value in practical alignment robustness scenarios.

This paper also analyzes the data performance of gated learning with the reduction of training sample size and label sparsity. The experimental results are shown in Figure 3.

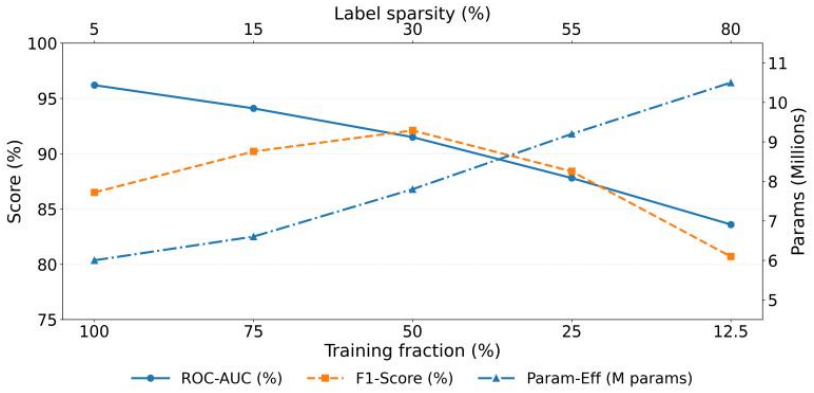


Figure 3. The impact of training sample size reduction and label sparsity on the data performance of gated learning

The results show that the ROC-AUC metric declines gradually as the training sample size decreases and label sparsity increases. This indicates that with insufficient supervision signals, the overall alignment robustness ability of the model is inevitably affected. However, the decline is relatively smooth, reflecting that the gating mechanism can still maintain robustness under highly sparse supervision. This demonstrates that the structured adaptation path proposed in this study can stably allocate limited parameter resources to support discrimination performance when data are insufficient.

The F1-Score exhibits a peak trend, being highest at medium training sample size and moderate sparsity, but lower at both extremes. Under these balanced conditions, the model achieves the best trade-off between precision and recall. When training data are abundant, performance declines due to overfitting and diminishing returns. When data are highly scarce, performance drops because of the imbalance in positive and negative samples under sparse supervision. This result shows that the gating module has an optimal point when regulating feature utilization. It can effectively suppress redundant updates and enhance alignment inconsistency capture.

The Param-Eff metric increases significantly as the sample size decreases and the label sparsity rises. This indicates that under insufficient supervision, the gating mechanism activates more trainable parameters to compensate for the lack of labels. This reflects the adaptive nature of structured adapters in resource utilization. They expand effective parameter channels to maintain modeling capacity for complex patterns. However, combined with performance metrics, it is clear that more parameters do not always bring continuous performance gains. This shows that a dynamic balance between efficiency and effectiveness is needed.

Taken together, the three metrics reveal the sensitivity of gated learning to data conditions. Under limited labels and reduced samples, the method achieves stable degradation instead of abrupt collapse by combining parameter adaptation and selective control of gating paths. The different trends of the metrics highlight the complementarity of the method in global alignment, fine-grained discrimination, and resource allocation. They also confirm the effectiveness of combining parameter-efficient fine-tuning with structured gating in data-constrained scenarios.

This paper also evaluates the environmental sensitivity of structured gating robustness under computational quota and memory constraints. The experimental results are shown in Figure 4.

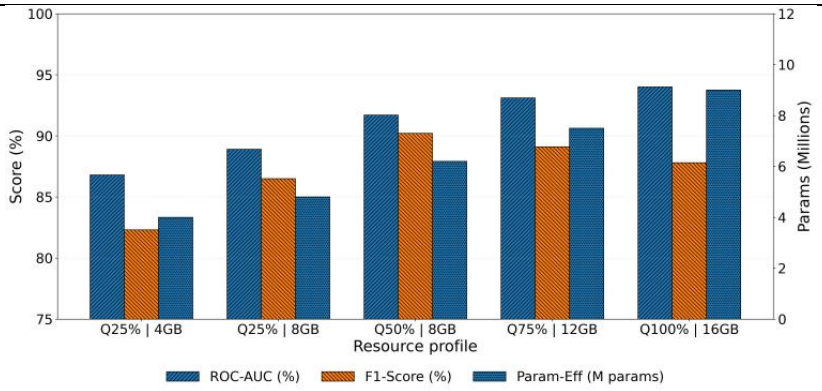


Figure 4. Environmental sensitivity of structured gating robustness under computational quota and memory constraints

The results show that the ROC-AUC metric increases monotonically with higher computation and memory resources. This indicates that the proposed structured gating method achieves stronger global representation and alignment robustness performance when more resources are available. The trend also suggests that although the gating mechanism can maintain basic performance under limited resources, its context modeling and temporal capture ability are further enhanced when computational and storage constraints are relaxed.

The F1-Score reaches a peak under medium to high resource conditions but decreases when resources are extremely low or extremely high. This peak-shaped variation reveals the sensitivity of the gating mechanism in feature utilization. When resources are insufficient, feature representation becomes limited, leading to lower recall. When resources are abundant, excessive parameter updates may introduce redundant information or overfitting risks, which disrupt the balance between precision and recall. This observation aligns closely with the idea of sparse adapter updates proposed in this study, highlighting that moderate computational and storage conditions are more favorable for stable model performance.

The Param-Eff metric increases steadily with more resources. This shows that under larger computational and memory budgets, the gating mechanism activates more trainable parameter channels to support complex feature modeling. The trend indicates that the method is highly adaptive, as it flexibly increases parameter utilization when resources improve. This enhances the depth and capacity of the model. Compared with traditional full fine-tuning, this gradual expansion ensures a balance between resource use and performance, which is consistent with the core goal of parameter-efficient fine-tuning.

Taken together, the three metrics demonstrate that the structured adapter gating mechanism exhibits differentiated adaptability under different resource conditions. It maintains basic alignment performance under low resources, achieves optimal balance under medium resources, and shows strong scalability under high resources. This flexibility not only enhances the applicability of the method in real deployments but also further verifies the robustness of combining parameter-efficient fine-tuning with gating in handling environmental sensitivity.

Next, this study conducted experiments on the environmental sensitivity of alignment stability under different hardware instruction sets and batch size settings. The experimental results are shown in Figure 5.

Under different hardware instruction sets and batch sizes, the ROC-AUC metric shows a rise followed by a decline. This indicates that when computational support is strong, the global representation ability of the model is better expressed, leading to higher alignment robustness accuracy. However, when the batch size becomes too large or the instruction set shifts to a lighter architecture, the model's ability to capture anomalies decreases. This suggests that the structured gating mechanism remains sensitive to uneven resource allocation, especially under high-throughput configurations where fine-grained alignment robustness evaluation is disrupted.

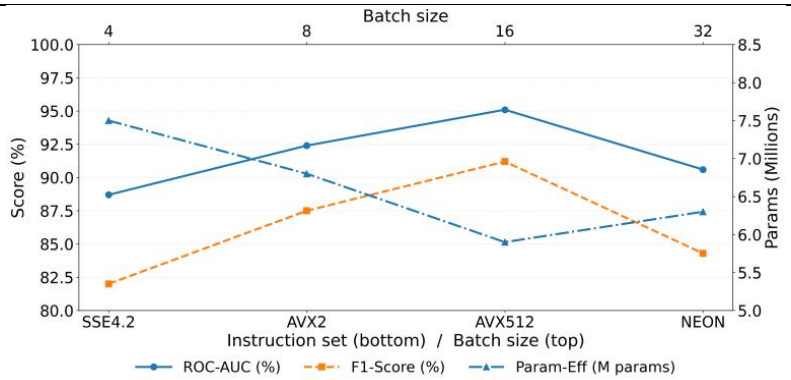


Figure 5. Alignment stability sensitivity under different hardware instruction sets and batch size settings

The F1-Score performs best under medium batch sizes and efficient instruction sets, but it is relatively weaker under very small or very large batch sizes. This reflects that under certain resource configurations, the model achieves a good balance between precision and recall. When resources are limited, recall declines. When parallelism is excessive, gradient noise or update conflicts may reduce precision. These results highlight the critical role of the gating paths in regulating feature utilization and suppressing redundancy.

The trend of Param-Eff is noticeably different. Resource utilization decreases under some instruction sets but rises again in lighter architectures. This indicates that the gating mechanism adapts parameter selection to the variation in instruction sets and batch sizes. It dynamically adjusts the scale of effective channels to handle different computational and memory pressures. This non-monotonic trend reflects the adaptive advantage of structured adapters in resource-constrained environments while also revealing their sensitivity to diverse hardware conditions.

Taken together, the three metrics show that the proposed parameter-efficient fine-tuning and structured gating method achieves both stability and flexibility under diverse hardware and batch configurations. ROC-AUC and F1 reflect changes in alignment quality and discrimination ability. Param-Eff reveals how the gating strategy reallocates parameter resources in different environments. The results indicate that the method maintains efficiency while adapting to the constraints of varying hardware and training settings.

Finally, this study analyzes the alignment robustness data under semantic noise and conflict feedback injection, and the experimental results are shown in Figure 6.

As the level of semantic noise increases, the ROC-AUC metric shows a continuous decline. This indicates that the structured gating method proposed in this study is affected in terms of alignment robustness and consistency ability when facing interfering information. The decline is expected, since strong noise weakens the consistency of feature signals and context, making it difficult for the model to maintain stable global judgments. However, the curve is relatively smooth, showing that the gating mechanism can partially resist noise disturbance, which reflects its robustness.

The F1-Score reaches a peak under moderate noise levels but is relatively lower under both low and high noise conditions. This “middle-optimal” pattern suggests that moderate perturbation can help the gating structure avoid overfitting and improve generalization. When noise becomes excessive, however, the balance between precision and recall is disrupted, leading to a clear performance drop. This phenomenon indicates that there is an optimal interval in the dynamic feature selection process of gating. Within this interval, effective information is maximized, and redundant feature interference is reduced.

The Param-Eff metric increases gradually with higher conflict feedback rates. This indicates that under stronger feedback conflicts, the gating module activates more adapter parameter channels to maintain overall performance. This trend reveals the adaptive nature of structured adapters. When task signals are unstable, the system increases parameter usage to offset uncertainty, thereby maintaining stable outputs in complex

feedback environments. Compared with traditional methods, the gating mechanism dynamically allocates parameter resources according to the environment, showing strong flexibility.

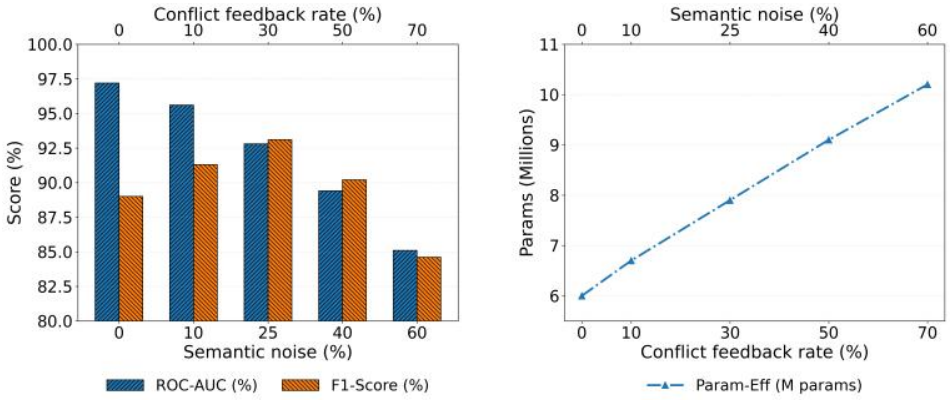


Figure 6. Data visualization of alignment robustness under semantic noise and conflicting feedback injection

By combining the three metric trends, it can be seen that under the joint influence of semantic noise and conflict feedback, the robustness of alignment shows both inevitable performance degradation and the adaptive compensatory capacity of the gating structure. The gradual decline of ROC-AUC, the peak change of F1-Score, and the steady increase of Param-Eff together indicate that parameter-efficient fine-tuning combined with structured gating can maintain stability under multi-source disturbances. At the same time, it dynamically balances performance and resource utilization, which represents the core value of the proposed method in alignment tasks.

5. Conclusion

This study focuses on parameter-efficient fine-tuning and structured adapter gating, and proposes a large model method that achieves high-quality alignment even under resource constraints and complex environments. By introducing sparse updates and gating control mechanisms, the model maintains low parameter cost while showing significant improvements in alignment robustness, stability, and controllability. The performance across different experimental settings further verifies the adaptability and stability of the method under hyperparameter sensitivity, data noise, and label sparsity, and computation and memory constraints. It provides a new perspective to address the common challenges of high cost and low portability in large model applications.

The experimental results show that the method achieves a balance between performance and efficiency across multiple evaluation metrics. It also maintains reasonable outputs in the presence of complex noise and conflicting feedback. This not only demonstrates the effectiveness of the proposed framework but also indicates that the model has strong adaptability when facing real-world challenges such as semantic perturbation, inconsistent feedback, and hardware heterogeneity. Such adaptability is especially important for large-scale distributed systems, cross-domain data analysis, and real-time monitoring. In these scenarios, environments are often dynamic and uncontrollable, and balancing stability with computational cost is a key issue for advancing the deployment of large models.

At the same time, the structured gating mechanism shows unique advantages in resource optimization. It can flexibly adjust parameter usage according to actual computational and storage conditions, avoiding the resource waste of traditional full fine-tuning. By dynamically expanding or shrinking adapter channels, the method enables the model to operate not only in high-performance hardware environments but also in resource-constrained edge devices. This provides strong technical support for deploying large models in areas such as intelligent manufacturing, financial risk control, medical diagnosis, and complex human – computer interaction, and it is expected to drive the wider adoption and deeper application of these technologies.

6. Future Work

Looking ahead, the combination of parameter-efficient fine-tuning and structured gating still offers broad research opportunities. Future work may study its applicability to cross-modal tasks, extending to multimodal fusion in vision, language, and speech. It may also explore its role in federated learning and privacy-preserving frameworks, providing low-cost and secure solutions for collaborative training. In addition, integrating dynamic scheduling strategies and more advanced adaptive mechanisms will be important to allow models to adjust online according to task demands and environmental states in real deployments. These directions will not only advance theoretical research but also enhance the practical value of large models across a wider range of applications.

References

[1] Han Z, Gao C, Liu J, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey[J]. arXiv preprint arXiv:2403.14608, 2024.

[2] Prottasha N J, Mahmud A, Sobuj M S I, et al. Parameter-efficient fine-tuning of large language models using semantic knowledge tuning[J]. Scientific Reports, 2024, 14(1): 30667.

[3] Z. Han, C. Gao, J. Liu, J. Zhang and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey", arXiv preprint arXiv:2403.14608, 2024.

[4] Z. Cheng, Y. Xu, M. Cheng, Y. Qiao, S. Pu, Y. Niu and F. Wu, "Refined Gate: A Simple and Effective Gating Mechanism for Recurrent Units", arXiv preprint arXiv:2002.11338, 2020.

[5] H. Inzirillo and R. Genet, "SigKAN: Signature-Weighted Kolmogorov-Arnold Networks for Time Series", arXiv preprint arXiv:2406.17890, 2024.

[6] A. Kumar, S. Garg and S. Dutta, "Uncertainty-Aware Deep Neural Representations for Visual Analysis of Vector Field Data", IEEE Transactions on Visualization and Computer Graphics, 2024.

[7] Y. Tu, B. Zhang, L. Liu, Y. Li, J. Zhang, Y. Wang et al., "Self-Supervised Feature Adaptation for 3D Industrial Anomaly Detection", Proceedings of the European Conference on Computer Vision, pp. 75-91, Sept. 2024.

[8] J. Gao, C. He, L. Duan and J. Zuo, "Towards Better Zero-Shot Anomaly Detection Under Distribution Shift With CLIP", Proceedings of the 35th British Machine Vision Conference (BMVC), 2024.

[9] Y. Cao, L. Lin and J. Chen, "Adversarially Robust Industrial Anomaly Detection Through Diffusion Model", arXiv preprint arXiv:2408.04839, 2024.

[10] Zamanzadeh Darban Z, Webb G I, Pan S, et al. Deep learning for time series anomaly detection: A survey[J]. ACM Computing Surveys, 2024, 57(1): 1-42.

[11] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng and Y. Jin, "Deep Industrial Image Anomaly Detection: A Survey", Machine Intelligence Research, vol. 21, no. 1, pp. 104-135, 2024.

[12] X. Ma, F. Liu, J. Wu, J. Yang, S. Xue and Q. Z. Sheng, "Rethinking Unsupervised Graph Anomaly Detection With Deep Learning: Residuals and Objectives", IEEE Transactions on Knowledge and Data Engineering, 2024.

[13] Audibert J, Michiardi P, Guyard F, et al. Usad: Unsupervised anomaly detection on multivariate time series[C]//Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 3395-3404.

[14] Tuli S, Casale G, Jennings N R. Tranad: Deep transformer networks for anomaly detection in multivariate time series data[J]. arXiv preprint arXiv:2201.07284, 2022.

[15] Deecke L, Ruff L, Vandermeulen R A, et al. Deep anomaly detection by residual adaptation[J]. arXiv preprint arXiv:2010.02310, 2020.

[16] Liu Y, Hu T, Zhang H, et al. itransformer: Inverted transformers are effective for time series forecasting[J]. arXiv preprint arXiv:2310.06625, 2023.