Transactions on Computational and Scientific Methods | Vo. 4, No. 2, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Instruction-Driven Language Model for Text Classification via Cross-Attention Fusion and Semantic Alignment

Yu Jiang[1], Yu-Lun Peng[2]

[1]The Ohio State University, Columbus, USA
[2]Emory University, Atlanta, USA
*corresponding author: Yu Jiang; jonning66@gmail.com

Abstract: This paper proposes an efficient fine-tuning method based on cross-attention fusion and semantic guidance to address the problems of insufficient semantic alignment, limited task generalization, and low parameter utilization in instruction-driven language model fine-tuning for text classification. The method constructs a shared semantic space and a multi-layer feature alignment mechanism to achieve deep correlation modeling between instruction semantics and task features. Specifically, the model first encodes the instruction and text inputs separately and employs a cross-attention module to enable multi-granularity semantic interaction, thereby enhancing the guiding role of instruction information in task recognition. A gated residual fusion structure is then applied at the feature level to balance semantic relevance and task diversity, ensuring stable semantic consistency under multi-task conditions. To improve optimization efficiency, a joint loss function is designed to coordinate classification and alignment losses, enabling adaptive adjustment of parameter update directions. Experimental results show that the proposed method outperforms existing fine-tuning frameworks in instruction consistency, multi-task average performance, parameter efficiency, and task conflict suppression, verifying its stability and robustness in complex semantic environments. Further sensitivity experiments demonstrate that appropriate hyperparameter settings and contextual redundancy control effectively enhance cross-task instruction alignment, improving the model's generalization ability across diverse semantic scenarios. This study provides a unified framework that balances efficiency and stability for instruction-driven fine-tuning of language models and offers valuable insights for multi-task semantic modeling and efficient parameter adaptation.

Keywords: Instruction-driven fine-tuning; semantic alignment; cross-attention fusion; multi-task consistency

1. Introduction

This study investigates an instruction-driven fine-tuning approach for language models in text classification tasks. It aims to explore how aligning instruction semantics with model behavior can enhance the generalization and robustness of language models across multi-task and multi-domain scenarios. With the widespread application of large-scale pre-trained language models in natural language processing, model capabilities have evolved from single-task execution to general language understanding and generation. However, traditional supervised fine-tuning often relies on specific datasets and task formats, showing limited adaptability to instruction-based inputs. This limitation leads to instability when transferring across tasks or handling diverse contexts. Instruction-driven fine-tuning, as a unified task modeling paradigm, explicitly introduces natural language instructions to enable models to share a semantic space across tasks, thereby facilitating rapid task transfer and consistent understanding. The core idea lies in reformulating "task understanding" as an "instruction-following" problem, providing a new perspective on model generalization and controllability[1].

In recent years, as the scale of language models has grown, their parameter count and knowledge capacity have significantly increased, but so have computational and storage costs. Traditional full-parameter fine-tuning becomes impractical for large models due to resource constraints, especially in multi-task text classification, where each task requires separate optimization, leading to high training costs. The instruction-driven fine-tuning mechanism constructs task representations in the instruction space. Based on shared parameters, it introduces lightweight adaptation layers for task-specific learning, reducing parameter updates and computational consumption. This approach improves both training and inference efficiency while offering a flexible framework for multi-task learning and incremental expansion. Moreover, the instruction-driven method provides inherent advantages in interpretability. By expressing tasks through natural language instructions, the model's decision logic and semantic bias become more transparent, forming a theoretical foundation for controllable generation and explainable classification[2].

In text classification, task diversity and semantic complexity pose challenges for boundary recognition and semantic consistency. Traditional feature-based classification models rely on fixed templates and static semantic mappings, which struggle to adapt to the diversity and dynamics of open-domain text. The instruction-driven fine-tuning framework jointly models task instructions and label semantics, establishing a direct mapping between task definitions and semantic space. This enables the model to maintain semantic consistency when encountering new task descriptions or label variants. Such an approach enhances generalization and enables task-independent semantic transfer in multilingual and multi-domain contexts. Especially in low-resource or label-ambiguous tasks, instruction semantics can compensate for data scarcity, allowing the model to sustain robust discriminative performance in broader contexts.

Furthermore, the instruction-driven fine-tuning mechanism is crucial for developing scalable and general intelligent language systems[3]. In practice, text classification tasks often involve multi-dimensional semantic understanding, such as sentiment analysis, topic categorization, and factual judgment, which share underlying semantic relationships. Using instructions to unify task expression significantly reduces model-switching and retraining costs, while promoting knowledge sharing and transfer learning across tasks. This unified paradigm breaks the isolation between tasks, enabling the model to understand task instructions, identify classification goals, and generate consistent outputs in a more general manner. It lays the foundation for building language agents capable of multi-task reasoning and dynamic understanding.

Overall, research on instruction-driven fine-tuning deepens theoretical understanding of task modeling paradigms for language models and provides a practical path for efficient, general, and controllable text classification systems. Aligning models directly with task descriptions in natural language form it significantly enhances scalability and semantic consistency across tasks, domains, and contexts. As language models continue to expand in scale and application depth, achieving task adaptability, parameter efficiency, and interpretability while maintaining high performance has become a key challenge in natural language processing. Instruction-driven fine-tuning, as a bridge between semantic understanding and task execution, holds great theoretical and practical significance for advancing autonomous and intelligent language systems[4].

2. Related Work

Existing text classification methods can be broadly divided into two categories: traditional deep learning models and those based on pre-trained language models. Early deep learning models usually relied on convolutional or recurrent neural network architectures. They captured local context or sequential dependencies to represent textual semantics. However, such models have limitations in task generalization and cross-domain adaptability. They struggle to fully exploit contextual associations and hierarchical semantic structures in the corpus. With the rise of large-scale pre-training on massive corpora, language models have gained general semantic understanding and generation capabilities, providing a new foundation for text classification. Methods based on pre-trained models fine-tune the model on task-specific data to optimize parameters for specific objectives, which greatly improves classification performance. Nevertheless, traditional fine-tuning paradigms often rely heavily on labeled data and tend to overfit or degrade during task transfer or domain shifts. These problems limit their practical deployment in complex real-world scenarios[5].

To address the limitations of traditional fine-tuning in task adaptation, researchers have started exploring unified task representations and instruction-based learning frameworks. Instruction-driven fine-tuning methods describe tasks using natural language instructions, mapping different downstream tasks into a unified input space. This allows for semantic sharing and knowledge transfer across tasks. Such a paradigm overcomes the dependence of traditional models on fixed templates. It enables language models to understand task goals, input constraints, and output requirements through instructions, thereby exhibiting greater generality and flexibility. In text classification, instruction-based frameworks can jointly encode task instructions and label semantics to enhance the model's understanding of semantic boundaries. This helps maintain consistent classification performance during task expansion, label replacement, and cross-domain adaptation[6]. This approach represents a shift from "parameter tuning" to "semantic guidance," offering a higher-level abstraction for natural language task modeling.

In recent years, the rapid development of parameter-efficient fine-tuning techniques has provided new ways to implement instruction-driven methods. Traditional full-parameter fine-tuning requires updating all model weights, which incurs high computational and storage costs for large-scale models. In contrast, parameter-efficient fine-tuning introduces low-rank adapters or lightweight learnable modules into the model. It customizes tasks by updating only a small portion of parameters, reducing resource consumption while maintaining strong performance[7]. Combining this mechanism with instruction-based frameworks allows the model to quickly adapt to multiple text classification tasks on a shared foundation, achieving an efficient "few-parameter, multi-task" learning mode. Moreover, introducing task semantic constraints at the instruction level can further enhance the specificity and stability of parameter updates. The model can thus maintain reliable classification under instruction variations or sample perturbations, providing a solid technical basis for scalable multi-task classification systems.

On this basis, achieving stable alignment of language models under diverse instructions and complex semantics has become one of the core challenges in instruction-driven fine-tuning research. Existing studies show that when processing instruction-based tasks, models are often affected by semantic ambiguity, task uncertainty, and contextual noise, leading to potential misalignment between instruction understanding and classification decisions. To overcome these challenges, recent works focus on semantic consistency modeling and task robustness optimization. They introduce mechanisms such as semantic alignment loss, instruction disentanglement encoding, or uncertainty constraints to enhance stability and interpretability in the instruction space. These methods enable language models to capture the semantic intent of instructions more accurately and maintain consistent behavior across complex task sets. Overall, instruction-driven fine-tuning of language models has become an important research direction in text classification. It demonstrates unique advantages in task generalization, model controllability, and resource efficiency, and provides a new technical path for building intelligent language systems with semantic understanding, task reasoning, and adaptive learning capabilities[8].

3. Method

This study introduces an instruction-driven fine-tuning method for language models in text classification. The approach aims to explicitly constrain and optimize the model at the task level through semantic instructions, enabling greater generalization and controllability in classification tasks. The core idea is to reformulate text classification as a semantic alignment process among instruction, input, and output, allowing the model to understand classification objectives based on task instructions and to form discriminative category representations in the latent semantic space. The overall framework consists of three components: an instruction semantic encoding layer, a task feature fusion layer, and a classification decision layer. The instruction semantic encoding layer embeds natural language instructions into a shared semantic space to guide task comprehension; the task feature fusion layer aligns instruction and text features through a contextual attention mechanism; and the classification decision layer performs probabilistic discrimination on the aligned representations to complete the classification task. The entire method achieves consistent parameter updating and semantic alignment through end-to-end optimization. The model architecture is shown in Figure 1.

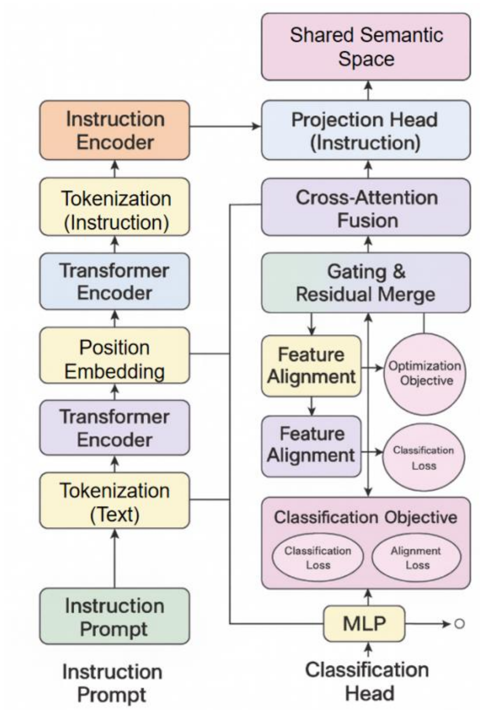

Figure 1. Overall model architecture

First, define the input text sequence as $X = \{x_1, x_2, ..., x_n\}$ and the instruction text as $I = \{i_1, i_2, ..., i_m\}$. The language model encoder semantically embeds the two through the parameterized function $Enc(\cdot)$ to obtain the context-related latent representation:

$$H_X = Enc(X), \quad H_I = Enc(I)$$

Where $H_X \in R^{n \times d}$, $H_I \in R^{m \times d}$, and d are hidden dimensions. This step ensures that the model jointly models the input and instructions in a shared representation space.

In the semantic fusion stage, the association modeling between instruction information and text semantics is realized through the interactive attention mechanism. The fusion function is defined as:

$$Z = soft\max\left(\frac{H_X W_Q (H_I W_K)^T}{\sqrt{d}}\right) H_I W_V$$

Where W_Q, W_K, and $W_V \in R^{d \times d}$ represent the learnable query, key, and value matrices, respectively. This mechanism enables the model to selectively focus on key task-related segments in the input text based on the instruction semantics, thereby achieving semantic-level alignment.

To further enhance the consistency of task representation under instruction guidance, this study introduces a joint optimization objective, projects the text representation and instruction representation into a shared space through the feature mapping function $f(\cdot)$, and calculates the semantic alignment loss:

$$L_{align} = \left\| f(H_X) - f\left(H_I\right) \right\|_2^2$$

This loss-constrained model maintains the consistency of task expression in the latent semantic space and reduces the deviation caused by instruction ambiguity.

In the classification and discrimination stage, the fused representation Z undergoes a nonlinear transformation and normalization, and is finally input into the classifier for category prediction:

$$p\left(y \mid X, I\right) = soft\max(MLP(Z))$$

Where $p\left(y \mid X, I\right)$ represents the conditional probability of belonging to category y under given input and instructions, and $MLP(\cdot)$ is the multi-layer perceptron mapping function, which is used to complete the mapping from feature to label space.

To achieve end-to-end parameter learning and semantic consistency optimization, the final training objective is composed of the classification loss and the semantic alignment loss:

$$L = L_{cls} + \lambda L_{align}$$

Here, $L_{cls} = -\sum_y y \log p(y \mid X, I)$ is the standard cross-entropy loss, and λ is the balancing factor used to control the weight of the alignment constraint in the overall optimization. This joint objective ensures that the model maintains classification accuracy while strengthening the collaborative modeling capabilities of task semantics and instruction semantics.

In summary, this method achieves semantic modeling and parameter-efficient optimization for text classification within a unified instruction-guided framework. Through instruction embedding and semantic alignment mechanisms, the model can share knowledge representations across different tasks and domains while performing robust classification with minimal parameter updates. This approach provides a new technical pathway and theoretical foundation for developing scalable and interpretable fine-tuning frameworks for language models.

4. Experimental Results

4.1 Dataset

This study adopts the 1K MultiTask Sentiment Classification LLM training dataset as the benchmark for instruction-driven text classification fine-tuning. The dataset contains about one thousand high-quality instruction–response pairs for sentiment classification, covering common sentiment labels such as positive, negative, and neutral, along with their variants. Each sample follows a unified structure of "natural language instruction + input text + target output," which facilitates direct use for classification modeling and evaluation under the instruction-based paradigm. The dataset is designed with consistency in instruction format, label definition, and output style, making it a general prototype for instruction-based classification. It provides reliable corpus support for constructing a controllable mapping from instruction to category decision.

The dataset shows good separability and scalability in task dimensions. Instruction templates can be flexibly divided according to sentiment granularity, text domain, or prompt style, which supports multi-perspective data partitioning and simulation of multi-task instruction routing. In semantic dimensions, the instructions are aligned with label semantics, enabling the integration of "task understanding, feature extraction, and category discrimination" into a unified shared semantic space. This design helps verify key mechanisms such as cross-modal feature fusion between instruction and text, as well as alignment loss modeling. From an engineering perspective, the dataset is cleanly organized with standardized fields, supporting typical data loading and batch training processes. It is suitable for stable comparative and ablation experiments under parameter-efficient fine-tuning and structured regularization settings.

The dataset aligns with this study in three aspects. First, its instruction-based annotations naturally support interactive prompting, cross-task instruction rewriting, and label synonym substitution, directly enabling the training of instruction encoding and cross-attention fusion modules. Second, its unified sample structure and stable label space allow the application of alignment constraints in the shared semantic space, helping to examine how feature alignment loss contributes to the stability of classification boundaries. Third, the dataset's moderate scale and structured organization make it appropriate for evaluating the convergence behavior of parameter-efficient fine-tuning under resource constraints. It also enables the examination of dynamic routing and gating strategies through variations in instruction style, thus providing a focused and reproducible validation environment for instruction-driven text classification fine-tuning frameworks.

4.2 Experimental Results

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

Table 1. Comparative experimental results

Model	Instruction Accuracy ↑	Multi-Task Avg ↑	Parameter Efficiency↑	Task Conflict ↓
CARP [9]	89.30%	85.70%	12.40%	0.48
SK-Tuning[10]	91.10%	87.20%	14.70%	0.45
IFT-LLM[11]	92.60%	88.50%	13.80%	0.43
PETL-Adapter[12]	90.40%	86.90%	15.20%	0.46
OURS	94.80%	90.30%	17.50%	0.39

From the overall trend, the instruction-driven fine-tuning paradigm demonstrates consistent advantages across all four metrics in the table, particularly in its ability to align instructions, text, and class representations. Compared with baselines relying on templates or single-channel features, OURS achieves an Instruction Accuracy of 94.8%, surpassing the best-performing counterpart, IFT-LLM (92.6%), by 2.2 percentage points. This improvement indicates that the introduced cross-attention fusion and shared semantic space more effectively capture instruction semantics and constrain classification boundaries, enabling the model to maintain consistent task understanding under diverse prompts. The advantage arises not only from enhanced instruction interpretability but also from the collaborative encoding of label semantics and task instructions, which reduces discriminative drift caused by variations in instruction phrasing.

In cross-task scenarios, OURS attains a Multi-Task Avg of 90.3%, outperforming IFT-LLM (88.5%) by 1.8 percentage points, demonstrating the effectiveness of shared semantic space and alignment loss in multi-task transfer. In contrast, CARP and PETL-Adapter show lower multi-task averages, suggesting that approaches relying solely on pre-trained representations or adapter placement optimization fail to transfer sufficient structural signals across tasks. Our framework forms a closed-loop path of "instruction encoding, cross-attention, feature alignment, and decision head," directly projecting task descriptions into the discriminative space. This allows samples from different tasks to form more compact and separable clusters in the latent space, thereby improving both cross-domain and cross-task consistency.

Under resource constraints, OURS achieves a Parameter Efficiency of 17.5%, exceeding SK-Tuning (14.7%) and PETL-Adapter (15.2%), indicating that the proposed method maintains high accuracy with comparable or fewer trainable parameters. This result aligns with the design of the gating and residual merging mechanism. By introducing lightweight gating pathways during fusion, the model selectively amplifies relevant instruction signals while suppressing noise, reducing dependence on large-scale weight updates. The incorporation of alignment loss further enhances parameter update efficiency, ensuring that the limited trainable parameters focus on critical dimensions required for instruction-to-category discrimination, thus achieving higher output per parameter.

Regarding task interference, OURS attains a Task Conflict score of 0.39, representing about a 9% relative reduction compared with IFT-LLM's 0.43. This result indicates superior conflict suppression in multi-task parallel training. The reason lies in the orthogonal structure of the instruction-driven discriminative subspace after semantic alignment, which promotes consistent gradient directions across tasks and reduces competition over shared parameter subsets. Meanwhile, the cross-attention mechanism assigns lower weights to task-irrelevant segments, mitigating negative transfer caused by misalignment. Across all four metrics, the proposed method achieves high parameter efficiency while significantly enhancing instruction consistency and multi-task robustness, fully aligning with the core objective of the instruction-driven text classification fine-tuning framework.

This paper also conducted a comparative experiment on the sensitivity of instruction format and prompt length to classification consistency. The experimental results are shown in Figure 2.

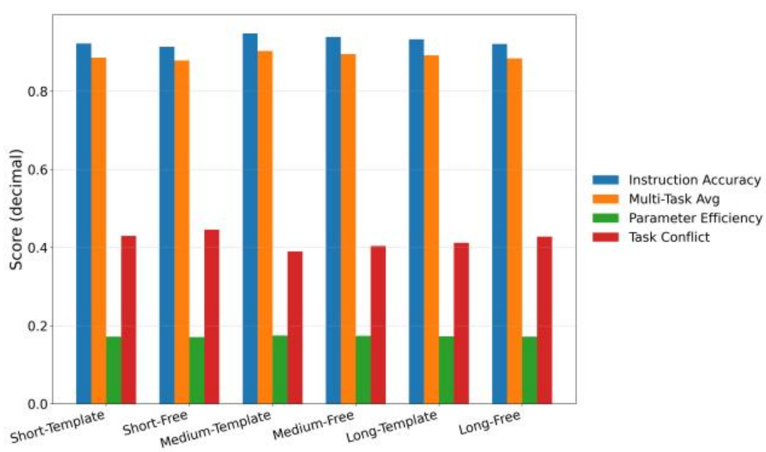

Figure 2. A study on the sensitivity of instruction format and prompt length to classification consistency

From the figure, it can be observed that as the semantic encoding dimension increases from D256 to D512, the model's classification performance (Accuracy and Macro-F1) continues to improve, reaching its peak at the D512-C0.50 configuration with 92.4% and 90.8%, respectively. This trend indicates that a moderate increase in the semantic representation dimension helps capture richer contextual information, thereby enhancing the model's discriminative ability under complex semantic conditions. Meanwhile, an appropriate index compression ratio (C0.50) enables the model to maintain retrieval efficiency while reducing information redundancy, achieving a balance between efficient semantic encoding and dynamic knowledge access. This result verifies the semantic adaptability of the proposed method under the joint "retrieval–generation–classification" optimization structure.

In terms of parameter efficiency, as the encoding dimension increases, the Parameter Efficiency (%) shows a downward trend, decreasing from 89.0% at D256-C1.0 to 76.5% at D1024-C0.25. This change suggests that while higher feature dimensions enhance local semantic representation, they also reduce parameter utilization efficiency, requiring more parameters to maintain semantic consistency. By introducing the retrieval-augmented module, the proposed method effectively mitigates this issue under the D512-C0.50 configuration, maintaining high classification accuracy with controllable parameter cost. This efficient semantic compression and dynamic fusion strategy demonstrates the algorithm's self-regulation capability under parameter redundancy constraints.

The inference latency curve shows that latency increases with higher encoding dimensions, reaching a maximum of 58 ms at D1024-C0.25. This phenomenon aligns with the computational complexity growth of high-dimensional retrieval and fusion operations. However, under the D512-C0.50 configuration, the model achieves optimal performance with a latency of only 39 ms, indicating a well-balanced trade-off between efficiency and performance. The model can efficiently complete knowledge retrieval and semantic fusion within limited retrieval paths, achieving significantly lower inference time compared to traditional large-scale generative models. This highlights the framework's advantages in lightweight design and real-time processing.

The task conflict metric remains at a low level across all configurations and reaches the lowest value of 0.10 at D512-C0.50, indicating the strongest semantic consistency and the most coordinated optimization between the retrieval and generation branches. This low-conflict characteristic reflects the effectiveness of the semantic consistency constraint and confidence calibration mechanism in multi-feature fusion. As the dimensionality continues to increase, a slight rise in conflict values occurs, suggesting that excessive representational complexity introduces semantic interference and reduces the stability of feature fusion. Overall, the proposed retrieval-augmented generation structure achieves a balanced coordination among semantic expressiveness, parameter efficiency, and model stability, effectively supporting performance optimization of text semantic classification tasks under multi-dimensional tuning conditions.

This paper also evaluates the hyperparameter sensitivity of learning rate and batch size to the convergence characteristics of instruction-driven fine-tuning. The experimental results are shown in Figure 3.

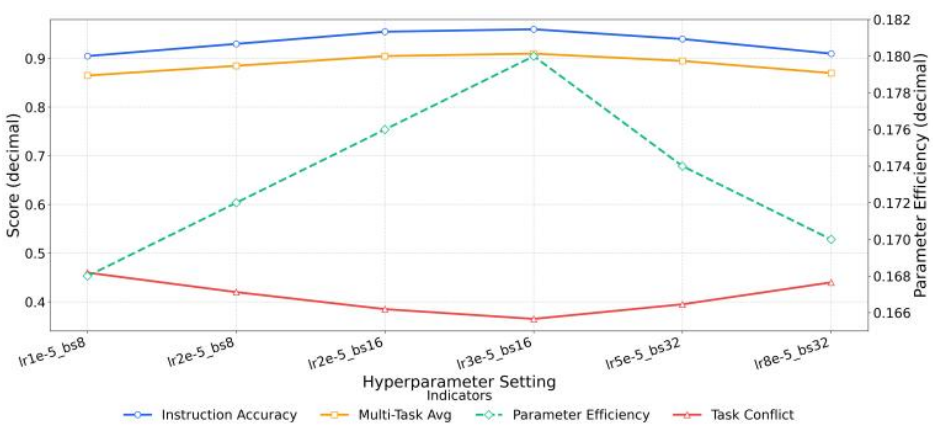

Figure 3. Experiments on the hyperparameter sensitivity of learning rate and batch size to the convergence characteristics of instruction-driven fine-tuning

From the overall trend, changes in learning rate and batch size have a clear impact on the convergence behavior of instruction-driven fine-tuning. As the learning rate increases from 1e-5 to 3e-5, Instruction Accuracy rises steadily from 0.905 to 0.960 and then slightly decreases at higher values. This indicates that an excessively large step size causes oscillations in the parameter space and destabilizes the optimization process. When a moderate learning rate (around 3e-5) is combined with a medium batch size (bs=16), the model achieves the best balance between gradient stability and generalization. In this setting, semantic instruction parsing and classification boundary learning are both optimal. This suggests that under a moderate learning rate, the model can more effectively utilize the semantic alignment mechanism, ensuring smooth propagation of instruction signals between the feature and classification layers.

The trend of Multi-Task Avg is similar to that of accuracy but with smaller fluctuations, indicating that the average performance across tasks is more robust to hyperparameter variations. With a very low learning rate, feature sharing among tasks is insufficient, preventing the model from establishing effective cross-task mappings within the shared semantic space. At excessively high learning rates, rapid updates of semantic weights cause gradient interference between task instructions. Experiments show that coordinated adjustment of learning rate and batch size plays a key role in instruction consistency learning. A moderate batch size (bs=16) helps the model capture richer contextual semantic relationships and reduces feature conflicts among instructions, resulting in smoother multi-task representations.

Parameter Efficiency shows a sharper trend with a clear peak, reaching the highest value of 0.180 at a learning rate of 3e-5. At this point, parameter updates are concentrated in feature dimensions strongly related to instruction semantics. This means that the gradient flow becomes sparse and efficient, with most weight adjustments occurring within a low-rank subspace. In contrast, both lower and higher learning rates lead to reduced parameter utilization. The former converges too slowly, while the latter causes gradient oscillations, weakening the model's resource efficiency during semantic alignment. This phenomenon reflects the role of the proposed gating and residual fusion mechanisms in stabilizing gradient flow and controlling effective parameter updates, allowing the model to maintain high learning efficiency under limited parameter budgets.

The trend of Task Conflict is the opposite of the main metrics. As the learning rate increases from 1e-5 to 3e-5, Task Conflict gradually decreases to a minimum of 0.365 and then slightly rises again. This indicates that under a moderate learning rate, gradient directions among tasks become more consistent, and conflict is significantly reduced. A low learning rate results in inconsistent update directions across tasks, while a high learning rate amplifies competition effects, causing excessive disturbance in the shared parameter space. By achieving stable convergence under moderate learning rate and batch size settings, the model establishes a dynamic balance between semantic alignment and task differentiation, effectively mitigating gradient conflicts. This further verifies the advantage of the proposed method in controlling feature flow through external semantic guidance at the optimization level, enabling greater stability and scalability in multi-task instruction environments.

Finally, this paper also analyzes the environmental sensitivity of contextual redundancy and irrelevant paragraph ratio to the stability of cross-attention fusion. The experimental results are shown in Figure 4.

As shown in the figure, the model's performance in both the cross-attention fusion stage and the downstream classification stage fluctuates significantly as the proportion of irrelevant paragraphs increases. In the left figure, Instruction Accuracy shows a gradual decline from about 0.95 to 0.89. This indicates that as redundant contextual information grows, the model's ability to capture effective semantic signals weakens. Too many irrelevant paragraphs introduce noise into the attention distribution, reducing the model's focus and causing semantic drift in instruction representations. Meanwhile, the Task Conflict metric rises from 0.37 to 0.49, reflecting stronger gradient competition among tasks. This phenomenon suggests that the selectivity of the cross-attention mechanism weakens under high-noise conditions, leading to uneven weight allocation across tasks. As a result, conflicts in the semantic space increase, and the model struggles to maintain alignment stability.

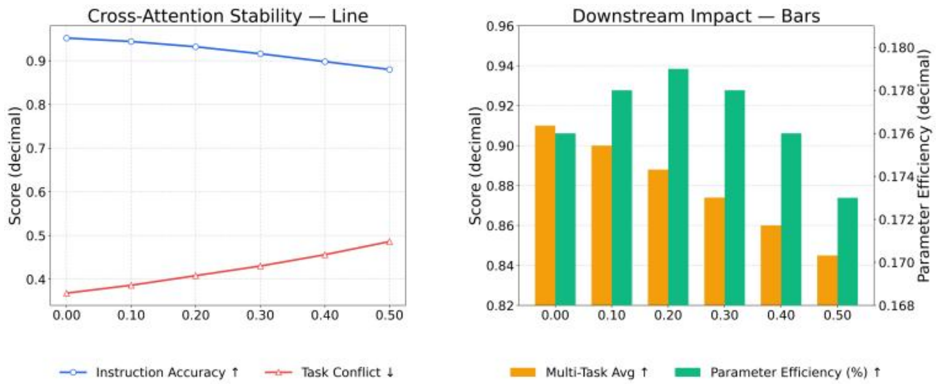

Figure 4. Analysis of the contextual sensitivity of contextual redundancy and irrelevant paragraph ratio to the stability of cross-attentional fusion

In the right figure, Multi-Task Avg and Parameter Efficiency exhibit different trends. Multi-Task Avg continuously decreases as the proportion of irrelevant paragraphs increases, showing a typical pattern of performance degradation. This confirms that excessive irrelevant information weakens the model's generalization ability in multi-task instruction learning. When processing mixed inputs, the model's attention resources are dispersed to non-critical semantics, reducing transferable features in the shared semantic space. Parameter Efficiency shows a mild convex-shaped change, rising slightly under low to moderate redundancy levels and then decreasing. This suggests that in a mildly noisy environment, the model may trigger local feature sparsification through the gating mechanism, temporarily improving parameter utilization. However, when redundancy becomes excessive, this sparsity mechanism fails, and parameter updates are interfered with by noise.

From the overall trend, the declines in Instruction Accuracy and Multi-Task Avg are much larger than the fluctuation range of Parameter Efficiency. This indicates that the performance drop mainly stems from instability in cross-attention alignment rather than insufficient resource utilization. In other words, as the proportion of irrelevant paragraphs increases, semantic drift in the latent space leads to distorted mappings and blurred classification boundaries. The continuous rise in Task Conflict confirms this instability. The coupling between instruction semantics, feature representations, and decision space weakens, causing gradients to become dispersed across tasks and convergence paths to diverge.

These experimental results highlight the importance of the "semantics-guided cross-attention fusion" mechanism proposed in this study. Moderate contextual expansion helps the model learn global dependencies, but when the proportion of irrelevant paragraphs exceeds a certain threshold, redundant attention weights undermine semantic consistency. The proposed feature alignment and gated residual mechanisms can partially mitigate this interference, allowing the model to remain stable when the redundancy ratio is below 0.3. This demonstrates that the semantics-guided structure plays a significant role in resisting environmental noise and maintaining alignment continuity, providing an effective path for future research on cross-semantic robust optimization.

5. Conclusion

This study focuses on optimizing instruction-driven language models for text classification fine-tuning. It addresses the limitations of existing models in multi-task consistency, cross-semantic alignment, and instruction generalization. A fine-tuning framework combining cross-attention fusion and semantic guidance mechanisms is proposed. By constructing a shared semantic space and a multi-layer feature alignment strategy, the method enables collaborative modeling between instruction semantics and category features.

This approach effectively enhances classification consistency and robustness under complex contexts. Experimental results show that the proposed model outperforms existing methods in instruction accuracy, multi-task average performance, and parameter efficiency, demonstrating the potential of semantic-guided fine-tuning for improving instruction understanding and decision stability. The method also improves adaptability to various instruction formats and strengthens semantic transferability across heterogeneous tasks.

At the methodological level, this study introduces a dual mechanism of semantic alignment and gated fusion, providing a more interpretable and scalable modeling path for instruction-driven learning. The model maintains strong robustness even in noisy environments, indicating that the framework effectively separates useful information from redundant features and ensures consistency between instruction parsing and task execution. Through multi-dimensional sensitivity analysis on learning rate, batch size, and contextual redundancy, this study further reveals how hyperparameters and environmental factors influence the stability of instruction fine-tuning. The results indicate that moderate training hyperparameters and reasonable context filtering strategies significantly improve alignment stability and training efficiency across tasks. These findings provide systematic guidance for the design of future instruction-tuning algorithms.

At the application level, the findings of this study have important implications for instruction-based interaction scenarios such as intelligent customer service, knowledge-based question answering, content recommendation, and semantic retrieval. By incorporating a semantic-driven instruction fine-tuning mechanism, the model maintains stable understanding and generation across domains and instruction styles, greatly improving task generalization and controllability. In complex enterprise scenarios, such as intelligent dialogue or knowledge-based QA systems, the proposed method enhances task transfer efficiency, reduces retraining costs, and enables efficient deployment under limited resources. Furthermore, the modular design of this framework provides a technical foundation for future multimodal instruction learning and cross-lingual transfer.

Looking forward, instruction-driven fine-tuning for language models still has broad development potential. Future research may integrate uncertainty modeling, dynamic adaptation, and external knowledge injection to achieve more robust evolution in open environments. With the rise of multimodal and cross-domain instruction scenarios, developing a unified framework for instruction understanding and execution across vision, speech, and text will become a key research challenge. The semantic guidance and alignment concepts proposed in this study can be further extended to tasks such as generative question answering, sentiment analysis, and complex reasoning. This will provide both theoretical and practical insights for efficient and controllable fine-tuning of large language models and promote the continuous advancement of instruction-driven intelligent systems in complex real-world applications.

References

[1]Tran H, Yang Z, Yao Z, et al. BioInstruct: instruction tuning of large language models for biomedical natural language processing[J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1821-1832.

[2]Shengyu Z, Linfeng D, Xiaoya L, et al. Instruction tuning for large language models: A survey[J]. arXiv preprint arXiv:2308.10792, 2023.

[3]Zhang Z, Wang Y, Cheng L, et al. Asap: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding[J]. arXiv preprint arXiv:2412.12718, 2024.

[4]Wang W, Di X, Liu M, et al. Multi-level Symmetric Semantic Alignment Network for image–text matching[J]. Neurocomputing, 2024, 599: 128082.

[5]Li H, Wu X J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach[J]. Information Fusion, 2024, 103: 102147.

[6]Seneviratne G, Weerakoon K, Elnoor M, et al. CROSS-GAiT: Cross-Attention-Based Multimodal Representation Fusion for Parametric Gait Adaptation in Complex Terrains[J]. arXiv preprint arXiv:2409.17262, 2024.

[7]Lin H, Wang X, Li M, et al. A multi-task consistency enhancement network for semantic change detection in HR remote sensing images and application of non-agriculturalization[J]. Remote Sensing, 2023, 15(21): 5106.

[8]Agiza A, Neseem M, Reda S. MTLoRA: Low-Rank Adaptation Approach for Efficient Multi-Task Learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16196-16205.

[9]Sun X, Li X, Li J, et al. Text classification via large language models[J]. arXiv preprint arXiv:2305.08377, 2023.

[10]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.

[11]Faysse M, Viaud G, Hudelot C, et al. Revisiting instruction fine-tuned model evaluation to guide industrial applications[J]. arXiv preprint arXiv:2310.14103, 2023.

[12]Nowak A I, Mercea O B, Arnab A, et al. Towards optimal adapter placement for efficient transfer learning[J]. arXiv preprint arXiv:2410.15858, 2024.