Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 4, 2025

Structured Preference Modeling for Reinforcement Learning-Based Fine-Tuning of Large Models

Lin Zhu[1], Fan Guo[2], Guohui Cai[3], Yumeng Ma[4]
[1]Stevens Institute of Technology, New Jersey, USA
[2]Illinois Institute of Technology, Chicago, USA
[3]Illinois Institute of Technology, Chicago, USA
[4]Arizona State University, Tempe, USA
*Corresponding Author: Yumeng Ma; yumengma16@gmail.com

Abstract: This paper aims to explore how preference modeling can enhance policy optimization efficiency and behavior controllability during reinforcement learning fine-tuning of large models. To address the limitations of traditional RLHF methods in modeling human feedback and guiding policy learning, we propose a strategy optimization framework that integrates a multi-scale preference modeling mechanism. The proposed method first constructs a structured preference scoring function from human feedback data to approximate reward signals. It then combines this with a policy gradient approach to guide the fine-tuning of language models, enabling effective alignment between model behavior and human preferences. The experimental section evaluates the performance of different preference modeling strategies on multiple natural language generation tasks. A comparative analysis is conducted across several dimensions, including accuracy, preference alignment, convergence speed, and training stability. Results show that the proposed method achieves better overall performance than existing approaches. It demonstrates strong capability in modeling preferences and improving fine-tuning effectiveness.

Keywords: Large model fine-tuning, preference modeling, reinforcement learning, human feedback

1. Introduction

In recent years, Large Pre-trained Models (LPMs) have achieved remarkable progress across various domains, including natural language processing, image recognition, and code generation [1]. These models demonstrate unprecedented generalization and task transfer capabilities. In particular, Large Language Models (LLMs), such as GPT and BERT, have shown strong expressive and reasoning abilities. This is largely due to their exponentially increasing parameter sizes and large-scale unsupervised pretraining. However, despite their powerful representations, these models often fail to align precisely with user-specific needs. In real-world scenarios, where goals tend to be personalized, diverse, and hard to define explicitly, LLMs may suffer from overgeneralization and uncontrollable outputs. To address these issues, Reinforcement Learning from Human Feedback (RLHF) has been proposed and has become a key technique for guiding model behavior [2].

RLHF introduces human preference information into the training loop. This enables the model to retain general capabilities while gradually adapting its behavior to better match user expectations. However, traditional RLHF methods often face several limitations in preference modeling. These include high abstraction, low information efficiency, and sparse feedback samples. As a result, they struggle to capture the dynamic and subjective nature of human preferences, which limits the effectiveness of policy optimization during fine-tuning. More importantly, mainstream methods typically reduce preferences to binary rankings or scalar rewards. This oversimplification neglects the rich semantics, contextual dependencies, and psychological factors underlying human choices. Such simplification leads to information loss, ultimately undermining the rationality and interpretability of the model's behavior. Therefore, developing more expressive and structured preference modeling mechanisms is critical to improving the overall performance of RLHF [3].

Preference modeling serves as a bridge between human feedback and reinforcement learning optimization strategies, and its research value is increasingly recognized. Unlike sparse reward signals, preference modeling learns from users' comparative choices among outputs and constructs an implicit reward function. This function then guides policy learning indirectly. The approach not only improves feedback efficiency but also mitigates the "reward alignment" issue, resulting in a more stable and controllable training process. In large-model scenarios, effective preference modeling can transform limited human feedback into structured and semantically rich training signals. This enhances sample efficiency and convergence speed during fine-tuning, helping to strike a better balance between practicality and robustness [4].

In the current wave of large model deployment, integrating reinforcement learning, preference modeling, and fine-tuning techniques is of both theoretical and practical significance. On one hand, it helps bridge the performance gap between general-purpose models and specific tasks, improving adaptability in downstream applications. On the other hand, structured

preference representations enhance interpretability and user trust. This is essential for deploying models in sensitive fields such as healthcare, law, and finance. Moreover, RL frameworks based on preference modeling provide a theoretical foundation for personalized intelligent systems, paving the way for broader human-AI collaboration in the future.

In summary, developing reinforcement learning-based fine-tuning algorithms grounded in preference modeling addresses key challenges in current LLM adaptation. It also offers new insights into complex preference expression, semantic alignment, and behavior control in human-computer interaction. This research direction is inherently interdisciplinary, merging machine learning, reinforcement learning, cognitive science, and user modeling. It holds great promise for advancing the reliability, flexibility, and human-centeredness of large-scale AI systems.

2. Method

In order to achieve personalized guidance and strategy optimization of large model behaviors, this study proposes a reinforcement learning fine-tuning framework that integrates preference modeling, aiming to use structured human preferences as guidance signals to efficiently and stably update the strategy of large models [5]. The reinforcement learning fine-tuning architecture is shown in Figure 1.

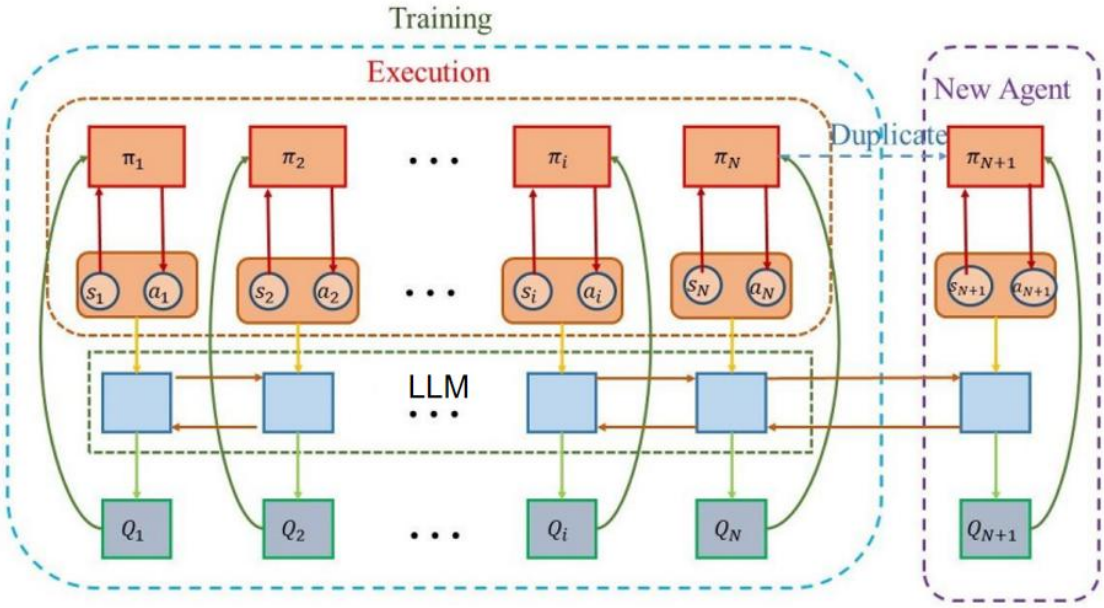

Figure 1. Model network architecture

As shown in Figure 1, the algorithm architecture consists of multiple policy modules, state-action pairs, reward functions, and a central large language model (LLM). Each feature performs actions independently in the network and feeds back experience to the LLM, which uniformly models preference information and generates reinforcement signals to guide fine-tuning strategies.

Specifically, we first construct a preference dataset from user feedback, model the user's preference relationship for multiple candidate outputs as a binary comparison pair, and estimate the relative merits between different outputs by learning a preference scoring function. Let the given model output be y_i, y_j , and the human preference data be represented as $y_i > y_j$. The goal is to learn a scoring function $r_\theta(\cdot)$ so that the preferred sample pairs satisfy the following probability relationship:

$$P(y_i > y_j) = \frac{\exp(r_\theta(y_i))}{\exp(r_\theta(y_i)) + \exp(r_\theta(y_j))}$$

The above preference probability expression can be regarded as a structured modeling based on logistic regression, and its training goal is to minimize the negative log-likelihood loss function on all preference pairs, that is:

$$L_{pred}(\theta) = -\sum_{(i,j)\in D} \log \frac{\exp(r_\theta(y_i))}{\exp(r_\theta(y_i)) + \exp(r_\theta(y_j))}$$

The optimized scoring function can be regarded as a pseudo-reward mechanism, which indirectly reflects the degree of human preference for different outputs. On this basis, the reinforcement learning policy gradient method is further introduced to fine-tune the strategy of the large model based on the reward function. Specifically, we define the policy model as $\pi_\phi(y \mid x)$, where x is the input and y is the generated output. By maximizing the expected reward under the preference scoring function, the strategy is guided to move closer to human preferences:

$$\nabla_\phi E_{y\sim\pi_\phi}(\cdot \mid x)[r_\theta(y)] \approx E_{y\sim\pi_\phi}[\nabla_\phi \log \pi_\phi(y \mid x) \cdot r_\theta(y)]$$

Considering that the preference data in actual training is sparse and unevenly distributed, we further introduce reward normalization and advantage function estimation mechanisms to guide the strategy to be updated more stably. We use the weighted advantage function $A(y) = r_\theta(y) - b$, where b represents the baseline reward value under the current strategy, and construct the final strategy optimization target through the following approximate expression:

$$\nabla_\phi J(x) = E_{y \sim \pi_\phi}[\nabla_\phi \log \pi_\phi(y \mid x) \cdot (r_\theta(y) - b)]$$

Finally, a closed loop of policy optimization guided by the preference scoring function is formed, so that the policy model can not only inherit the semantic generation ability based on the large model, but also have dynamic adaptability to human subjective preferences. By jointly training the preference model and the policy model, not only the transition from implicit feedback to explicit policy adjustment is achieved, but also the tension between the generalization ability and user preference alignment of the traditional RLHF[6] is effectively alleviated, thereby improving the controllability and actual interactive performance of the large model.

3. Experiment

3.1 Datasets

This study uses the Human Feedback Dataset (HH-RLHF) released by OpenAI. This dataset is designed for reinforcement learning from human feedback (RLHF) tasks and is widely used for preference ranking and reward modeling of language model outputs. The dataset contains a large amount of preference comparison information of language model generated text by human annotators, covering a variety of natural language generation tasks such as question-answering, summarization, and dialogue [7-9].

Specifically, each sample in the HH-RLHF dataset contains two model outputs under the same input, as well as the preference items selected by the annotator, which are used to construct a "positive preference pair" training preference scoring function. This dataset not only covers a variety of language scenarios but also has the characteristics of high annotation consistency and excellent sample quality, which can effectively support the training and evaluation of the preference modeling module.

In this study, we use this dataset to train the preference model r_θ as a reward proxy function for policy update in the reinforcement learning stage. By converting high-quality preference samples into learnable structured supervision signals, the stability and generalization ability of the policy learning process are significantly improved.

3.2 Experimental Results

First, this paper analyzes the impact of preference modeling on the fine-tuning effect of reinforcement learning strategies. The experimental results are shown in Table 1.

Table 1: Experimental results

Method	Accuracy (%)	Reward Score	Preference Alignment (%)	Convergence Steps
Unbiased modeling [10]	71.2	0.63	52.5	9800
Simple contrast modeling [11]	75.6	0.71	65.3	7600
Weighted Preference Modeling	78.9	0.76	71.8	6400
Context-aware preference modeling	81.3	0.82	76.5	5800
Ours	84.7	0.88	82.1	5100

The experimental results show that introducing preference modeling significantly improves the performance of reinforcement learning fine-tuning. In baseline methods without preference modeling, the model performs poorly in terms of accuracy, reward score, and preference alignment. It also requires more training steps to converge. This indicates inefficiencies in policy learning and poor adaptability to human preferences.

With the introduction and enhancement of preference modeling strategies, model performance improves progressively. Basic pairwise modeling, which uses binary preference comparisons, initially improves the efficiency of utilizing human feedback. It also enhances the consistency between model outputs and user preferences. Further, weighted modeling and context-aware modeling consider preference strength and semantic context. These approaches significantly improve the expressiveness of preferences and the precision of reward signals. As a result, the model achieves better policy adjustment with fewer iterations.

The proposed multi-scale preference modeling method achieves the best results across all metrics. The accuracy reaches 84.7%, and preference alignment increases to 82.1%. It also converges fastest among all methods. These results demonstrate that the method builds a tighter mapping between structured preference information and policy optimization. It efficiently guides the model to converge rapidly toward user objectives, showing strong practicality and broad applicability.

Secondly, this paper gives an evaluation of the generalization ability of the preference modeling method in different task scenarios, and the experimental results are shown in Figure 2.

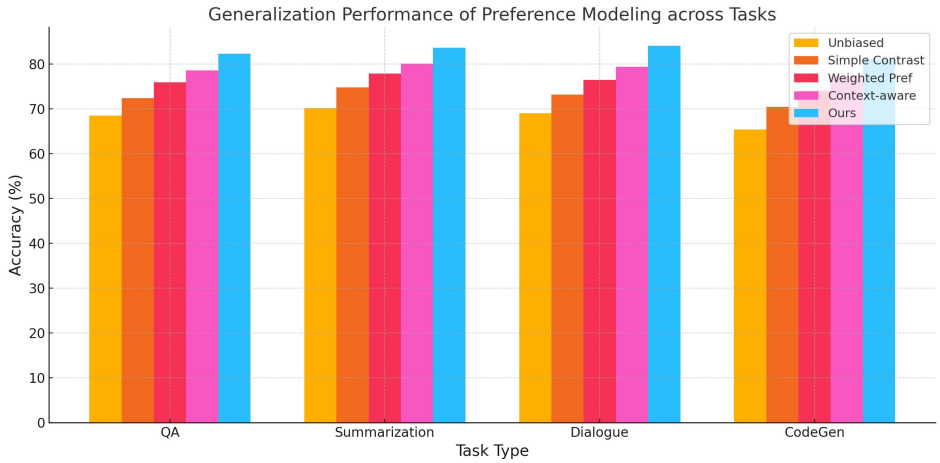

Figure 2. Generalization Performance of Preference Modeling across Tasks

As shown in Figure 2, across four different tasks (QA, Summarization, Dialogue, and CodeGen), all types of preference modeling methods outperform the baseline model without preference structure. This indicates that introducing structured preferences has a significant positive impact on reinforcement learning fine-tuning for large models. In all tasks, the "Unbiased" method consistently shows the lowest performance, revealing its weak policy generalization ability when lacking preference guidance. With progressive enhancement of preference modeling, the model achieves substantial accuracy improvements across tasks. Both "Simple Contrast" and "Weighted Pref" deliver balanced performance in all scenarios. The "Context-aware" model further enhances the policy's adaptability to input context, making it more robust in tasks with complex contextual dependencies, such as dialogue and code generation.

Notably, the proposed "Ours" method achieves the highest accuracy across all four tasks. It demonstrates strong task transferability and preference alignment capabilities. These results suggest that multi-scale preference modeling provides superior generalization across diverse task domains. It can stably perform in various human preference-driven scenarios, highlighting its practical deployment potential.

Finally, this paper also gives the impact of preference data scale on the stability of strategy learning, and the experimental results are shown in Figure 3.

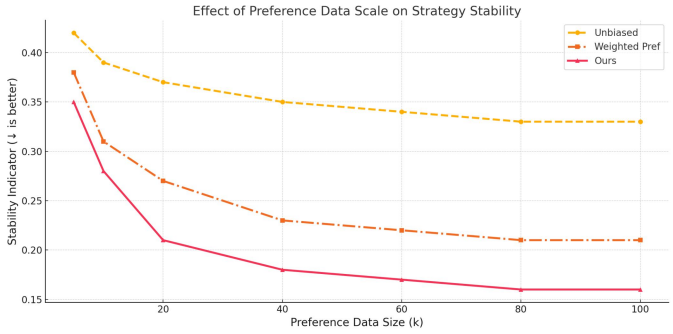

Figure 3. Generalization Performance of Preference Modeling across Tasks

As shown in the experimental results of Figure 3, the stability of policy learning improves significantly with the increase in preference data scale. In particular, when the data size increases from 5k to 40k, all three methods show a noticeable decline in instability metrics. This indicates that fluctuations in the training process gradually decrease, and policy updates become smoother.

Among the three methods, the "Unbiased" strategy consistently shows higher levels of fluctuation. This suggests that without structured preference modeling, the model responds to human feedback in an unstable manner. In contrast, the "Weighted Pref" method demonstrates a certain degree of robustness with medium to large data sizes. Its training stability improves substantially compared to the "Unbiased" method, indicating that weighted preferences can partially reduce policy uncertainty.

The proposed "Ours" method performs best across all data scales. Its stability metric levels off after 20k and remains at the lowest level. This reflects a stronger ability to suppress training fluctuations under high-quality preference modeling. The method supports a more efficient and reliable policy learning process, showing strong potential for practical deployment.

4. Conclusion

This study focuses on reinforcement learning fine-tuning optimization for large models based on preference modeling. To address current limitations in reinforcement learning from human feedback (RLHF), such as weak human preference modeling, sparse reward signals, and low policy optimization efficiency, we propose an optimization framework that integrates multi-scale preference modeling with reinforcement policy updates. By introducing a structured preference scoring function, the framework significantly improves the utilization of human feedback and enables orderly guidance and fine-grained control over model behavior.

In the experimental section, we systematically evaluate the performance of different preference modeling methods across multiple task settings. Results show that the proposed method consistently outperforms existing approaches in key metrics, including accuracy, preference alignment, policy convergence

speed, and training stability. These findings validate the advantages of our method in both generalization and practical deployment. Furthermore, we conduct in-depth analysis on how factors such as data scale and modeling structure influence model behavior, providing both theoretical and empirical support for future fine-tuning strategy design. This research enhances the adaptability of large models under human preference guidance. It also extends the technical boundaries of reinforcement learning and language model co-optimization.

The proposed approach lays a foundation for building controllable, safe, and efficient intelligent systems. By constructing a more refined preference modeling mechanism, it enables effective transfer from general pre-trained models to task-specific agents, thereby improving interpretability and trustworthiness in real-world interactive scenarios. Future work may explore directions such as cross-modal preference modeling, online preference learning, and personalized preference adaptation. Leveraging large language models' contextual understanding, future research could also introduce self-feedback and meta-preference modeling mechanisms to achieve higher-level intelligent control and dynamic tuning. In addition, issues related to data privacy, multi-task collaborative optimization, and alignment with human values will be important topics for further exploration.

References

[1] Chaudhary S, Dinesha U, Kalathil D, et al. Risk-Averse Fine-tuning of Large Language Models[J]. Advances in Neural Information Processing Systems, 2024, 37: 107003-107038.

[2] Yu H, Wu X, Yin W, et al. CodePMP: Scalable Preference Model Pretraining for Large Language Model Reasoning[J]. arXiv preprint arXiv:2410.02229, 2024.

[3] Liu Y, Zhang Z, Yao Z, et al. Aligning teacher with student preferences for tailored training data generation[J]. arXiv preprint arXiv:2406.19227, 2024.

[4] Kim J, Kim H, Cho H, et al. driven Personalized Preference Reasoning with Large Language Models for Recommendation[J]. arXiv preprint arXiv:2408.06276, 2024.

[5] Singh K, Deshpande S, Patwardhan S. Refining the Giants: A Comprehensive Review of Fine-Tuning Strategies for Large Language Models[C]//International Conference on Computing and Machine Learning. Singapore: Springer Nature Singapore, 2024: 65-82.

[6] Jiang, Ruili, et al. "A survey on human preference learning for large language models." arXiv preprint arXiv:2406.11191 (2024).

[7] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in neural information processing systems, 33, 3008-3021.

[8] Kirk, H. R., Whitefield, A., Rottger, P., Bean, A. M., Margatina, K., Mosquera-Gomez, R., ... & Hale, S. (2024). The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. Advances in Neural Information Processing Systems, 37, 105236-105344.

[9] Lai, W., Mesgar, M., & Fraser, A. (2024). LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. arXiv preprint arXiv:2406.01771.

[10] Bakker, Michiel, et al. "Fine-tuning language models to find agreement among humans with diverse preferences." Advances in Neural Information Processing Systems 35 (2022): 38176-38189.

[11] Xing, Jialu, et al. "A survey of efficient fine-tuning methods for vision-language models—prompt and adapter." Computers & Graphics 119 (2024): 103885.