Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA36530

# *Reinforcement Learning Methods and Optimization Strategies in Preference Alignment Techniques for Large Language Models*

**Hongyu Jiang**

*Management Information Systems, Beijing Jiaotong University (Weihai), Weihai, China*
*23711091@bjtu.edu.cn*

***Abstract.*** The growing pre-training scale has improved large language models' performance in language generation and knowledge representation. However, their training objectives remain limited to fitting data distributions, thus making it difficult to guarantee that the output satisfies human intentions and safety constraints. Therefore, the utilization of preference information to optimize model behavior and align generated content with human expectations has become a key challenge in the post-training phase of large language models. This paper investigates the development path of large language model preference alignment techniques, focusing on the training mechanism of Reinforcement Learning from Human Feedback (RLHF) and its three-stage process, including supervised fine-tuning, reward model training, and PPO-based policy optimization. On this basis, it compares the characteristics of Direct Preference Optimization (DPO), Reinforcement Learning from AI Feedback (RLAIF), and other methods in terms of training stability, data dependence, computational cost, and generalization ability. It further analyzes the alignment tax, which reflects the trade-off between improving model safety and preserving general capabilities during preference alignment. The results show that RLHF still possesses a higher upper bound for alignment in complex interactive scenarios, but its multi-stage training process brings high data and computational costs. In contrast, offline preference optimization methods such as DPO reduce training complexity and improve resource efficiency, but their performance remains constrained by preference data distribution and limited policy exploration capabilities.

***Keywords:*** Large Language Models, Preference Alignment, RLHF, Direct Preference Optimization (DPO), Alignment Tax

## 1. Introduction

Large Language Models learn language statistical patterns and contextual representations through the Next-token Prediction task on massive texts during the pre-training phase. However, self-supervised pre-training optimizes the objective of fitting text distributions and does not explicitly model human preferences and behavioral constraints. This may cause models to generate outputs that deviate from instructions, contain factual errors, or violate safety norms. To narrow the gap

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA36530

between model behavior and human expectations, alignment methods use feedback signals to optimize generation policies and guide models toward Helpful, Honest, and Harmless (HHH) responses [1]. Among these approaches, Reinforcement Learning from Human Feedback (RLHF) has been widely used for post-training large language models, with successful applications in dialogue models such as InstructGPT and ChatGPT [2]. However, the multi-stage training process and reinforcement learning optimization mechanism of RLHF bring high computational costs and training stability challenges. To reduce reliance on reward models and the online reinforcement learning process, non-reinforcement learning methods such as DPO have been proposed and are gradually becoming an important direction in preference optimization research. Existing review works have systematically analyzed the training mechanisms, performance, and application characteristics of different alignment methods [3, 4]. By reviewing existing research, this paper analyzes the mechanism by which RLHF translates human preferences into training signals, compares the improvements of DPO, RLAIF, and other methods on RLHF's training cost and stability issues, and discusses the trade-offs between safety, general capabilities, and data costs during the LLM alignment process.

## 2. Theoretical foundation and training mechanism of RLHF

### 2.1. Definition and problem modeling of RLHF

RLHF trains a reward model from human preferences to optimize language model policies via reward signals. This method inherits the idea of constructing optimization objectives through feedback signals in reinforcement learning, which is used to handle tasks where reward functions are difficult to define directly [5]. In natural language generation tasks, RLHF learns the relative quality of responses through preference comparison data, and further applies it to the post-training of large language models to align model behavior with human preferences [2, 6, 7].

In the RLHF framework, the text generation process is formulated as a sequential decision-making problem based on human preference feedback. For an input Prompt, the policy model $\pi_\theta$ generates a Token sequence step-by-step based on the current context, where the input context serves as the state and the generated response serves as the action. Since objectives such as text quality, factuality, and safety are difficult to explicitly define through fixed rules, RLHF introduces a Reward Model (RM) to evaluate response preferences and map human comparisons into computable scalar reward signals. The policy model updates its generation distribution using reward feedback, promoting preferred responses while constraining policy changes to preserve original capabilities.

### 2.2. Multi-stage training mechanism of RLHF

The PPO-based RLHF training process gradually converts human preferences into model parameter updates through three stages: "behavior initialization, preference modeling, and policy optimization." Among them, SFT establishes the model's instruction response capabilities, RM learns preference relationships between different outputs, and PPO further adjusts the model's generation policy based on reward signals.

In the SFT stage, the pre-trained model is fine-tuned using manually constructed Prompt-Response pairs, enabling it to learn the mapping between instructions and responses. As this stage relies merely on demonstration data for imitation learning, the model gains basic instruction-following capabilities but is unable to assess the relative quality of responses for the same input. Therefore, it is necessary to further introduce a reward model to translate human preferences into

DOI: 10.54254/2755-2721/2026.BA36530

optimizable evaluation signals. In the RM training stage, the model uses pairwise preference data to learn the relative superiority and inferiority relationships between responses. Due to the instability of absolute scoring across annotators, RLHF generally adopts pairwise preference comparison between a chosen response $y_w$ and a rejected response $y_l$ for the same prompt x. Given the reward model $r_\phi(x,y)$, the Bradley-Terry model defines the probability of human preference as:

$$P(y_w \succ y_l|x) = \sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right) \tag{1}$$

where $\sigma$ is the Sigmoid function, and the corresponding maximum likelihood loss is:

$$L_R(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)\right] \tag{2}$$

By minimizing this loss, the reward model learns response preferences and maps human judgments into scalar reward signals [2].

After obtaining the reward function, PPO is used to optimize the policy model $\pi_\theta$ to maximize the reward model score of its generated results. To prevent reward hacking caused by exploiting reward model flaws, PPO incorporates a KL divergence constraint to restrict policy deviations from the SFT reference model [8]. Its objective function is expressed as:

$$\text{Objective}(\theta) = \mathbb{E}_{(x,y)\sim D}\left[r_\phi(x, y) - \beta \mathbb{D}_{KL}\left(\pi_\theta(y|x) \,\|\, \pi_{SFT}(y|x)\right)\right] \tag{3}$$

where $\pi_{SFT}$ is the reference model obtained in the SFT phase, and $\beta$ is a hyperparameter controlling the strength of the KL penalty constraint. The term $\mathbb{E}_{(x,y)\sim D}$ represents the expectation calculated over a given prompt dataset $\mathbb{D}$, where x denotes the sampled input context and y means the response generated by the policy model. Since this process involves the collaborative updating of the policy model, reward model, value model, and reference model, PPO still has high requirements in terms of training stability and computational resources [9].

## 2.3. Key limitations and constraints of RLHF

RLHF optimizes language model generation policies via human preference signals, making the model output more aligned with instruction requirements and safety constraints [2]. However, its training process relies on high-quality preference data, reward models, and reinforcement learning optimization processes, so it is still limited by data costs, training complexity, and reward modeling errors [10].

More specifically, RLHF is constrained by the acquisition cost of preference data. This is because reward model training requires humans to rank responses. In professional domains such as mathematics, programming, and law, the accurate evaluation of model outputs often requires specialized knowledge, further increasing the difficulty of preference data construction. Meanwhile, the PPO stage requires the joint optimization of the policy model, reward model, value model, and reference model, and is highly sensitive to parameters like learning rate and KL constraint coefficients, leading to high computational overhead and debugging costs during the training process [9]. Furthermore, the reward model's ability to approximate human preferences limits the optimization effect of RLHF. Since the reward model learns from limited preference data, it may fail to capture the multi-dimensional nature of text quality. Reward model biases can cause the policy model to optimize superficial reward signals instead of improving true response quality. For

DOI: 10.54254/2755-2721/2026.BA36530

example, the model may obtain higher scores by adding redundant content or using pandering expressions, thus causing reward hacking [10]. Additionally, subjective tendencies in preference data may also be learned by the model and affect subsequent policy updates.

## 3. Large language model preference alignment methods based on feedback signals

### 3.1. Preference optimization methods based on human feedback

To address the multi-model collaborative training and optimization stability issues in the PPO stage of RLHF, subsequent research has focused on two aspects: preference data utilization and reinforcement learning improvement.

In the direction of direct preference optimization, Rafailov et al. proposed the Direct Preference Optimization (DPO) method [11]. Unlike PPO-based RLHF, DPO utilizes human preference data to directly update policy model parameters without explicitly training a reward model or performing the PPO online sampling process. Its core idea is to transform the reward optimization objective into a probability ratio optimization problem between the policy model and the reference model, achieving preference learning through implicit reward modeling. Its optimization objective is expressed as:

$$L_{DPO}\left(\pi_\theta;\pi_{SFT}\right)=-E_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{SFT}(y_w|x)}-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{SFT}(y_l|x)}\right)\right] \quad (4)$$

By removing the need for reward model training and PPO sampling procedures, DPO simplifies the training pipeline and improves optimization stability. On this basis, IPO further alleviates overfitting under limited preference data by modifying the optimization objective [12]. KTO remodels preference feedback based on prospect theory, enabling the model to optimize using non-paired data [13].

Unlike direct preference optimization, some studies maintain the reinforcement learning framework while improving the PPO optimization process. ReMax improves training efficiency by simplifying the reward optimization process, while RAFT utilizes a reward model to filter high-quality samples and approximates the reinforcement learning optimization process through supervised fine-tuning, reducing computational overhead [14, 15]. However, these methods still rely on reward model quality, and their effectiveness is limited by preference data coverage and reward modeling errors.

### 3.2. Alignment methods based on model feedback

To reduce the cost of acquiring manual preference data, alignment methods based on model feedback attempt to utilize the language model's own evaluation capabilities to generate preference signals. For example, Anthropic's Constitutional AI framework constrains model outputs through a predefined set of behavioral principles and enables the language model to self-critique and revise generated content according to these principles [16]. This method replaces manual safety evaluation with rule-guided model feedback, enabling more automated preference data generation.

Building upon Constitutional AI, RLAIF further incorporates AI feedback into the reinforcement learning alignment process. In particular, a teacher model compares multiple candidate responses and generates preference rankings, which are then used to train a reward model or directly optimize the policy model [17]. Through this process, RLAIF reduces human involvement, thus enabling more cost-effective scaling of preference data while retaining the feedback-driven policy

DOI: 10.54254/2755-2721/2026.BA36530

optimization paradigm. However, since preference judgments are made by the model, its feedback quality is influenced by the capabilities of the Teacher model and the consistency of its evaluation standards. Moreover, further research has begun to explore autonomous alignment methods that do not rely on external feedback models. Self-Reward Language Models introduce the LLM-as-a-Judge mechanism, enabling the model to simultaneously undertake response generation and quality evaluation tasks, and iteratively update training data using the generated preference signals [18]. This method reduces dependence on external annotations and separate evaluation models, enabling domain-specific alignment. However, shared capability limits between the generator and evaluator may reinforce errors and amplify biases. Thus, model-based feedback methods remain constrained by feedback reliability, evaluation consistency, and bias accumulation.

### 3.3. Comparison of methods driven by preference signals

Preference alignment methods differ in their feedback sources, optimization mechanisms, and training costs, as shown in Table 1. Specifically, RLHF uses manual preference data to train a reward model and combines PPO to update the model policy, thereby enabling model behavior adjustment through online optimization. In contrast, DPO directly optimizes the policy model based on preference data; by omitting reward model training and reinforcement learning sampling processes, it reduces training complexity. Based upon these approaches, RLAIF further changes the acquisition process of preference signals by reducing reliance on manual annotation and enabling large-scale preference data generation through model-based feedback.

Table 1. Comparison of mainstream LLM preference alignment techniques

| Technical Mechanism | Core Advantages | Main Disadvantages and Bottlenecks | Applicable Scenarios |
|---|---|---|---|
| Classic RLHF (PPO) [2, 7] | Online optimization, strong exploration, higher alignment potential | High computational cost, complex multi-model training, stability challenges [9] | Large-scale dialogue models and complex interactive tasks |
| Direct Preference Optimization (DPO) [11] | No reward model or RL sampling, simple and stable training | Dependence on offline preference data, limited exploration capability | Resource-constrained scenarios and rapid domain adaptation |
| RLAIF (AI Feedback) [16, 17] | Reduced annotation cost, scalable preference data generation | Dependence on feedback model quality, bias propagation risk | Low-cost alignment and automated preference data construction |

At the mechanism level, RLHF uses online policy optimization to retain exploration ability, but its training requires considerable computational resources and careful optimization stability management. In contrast, DPO improves efficiency through offline preference optimization but relies on preference data distribution. Meanwhile, RLAIF enhances preference data acquisition efficiency, but the quality of feedback still depends on the capability of the feedback-generating model. Thus, alignment methods balance performance, efficiency, and data cost.

DOI: 10.54254/2755-2721/2026.BA36530

## 4. Trade-offs between model capability and data dependence in RLHF alignment

### 4.1. Capability trade-offs and alignment impacts

Through preference alignment, the original capability distribution of the model is modified, resulting in improved safety and reduced performance in certain general capabilities, a phenomenon known as the "Alignment Tax." For example, evaluations of open-source models like Llama 2 and Gemma show that alignment training can affect performance in complex reasoning, code generation, and comprehensive knowledge tasks [19, 20]. This phenomenon results from shifts in the output distribution of the model during preference optimization. In the pre-training phase, the model fits a broad text distribution, while in the alignment phase, it adjusts response probabilities according to human preferences, strengthening outputs that meet evaluation standards and suppressing low-preference outputs. Reward model biases or limited preference data coverage may further undervalue reasonable but atypical responses, reducing the use of long-tail knowledge and complex reasoning strategies and degrading general capabilities.

Furthermore, preference optimization alters how the original capabilities of a model are transferred across different tasks. Model parameter updates affect not only safety-related behaviors but also the knowledge representations and generation patterns formed during the pre-training phase. When the optimization objective is overly focused on satisfying preferences, the model may tend to generate responses that are more compliant with evaluation standards at the expense of reasoning flexibility in some open scenarios. Different alignment methods exhibit varying degrees of capability retention due to differences in optimization objectives and parameter update methods. For example, DPO changes the generation distribution by adjusting the probability relationship between the policy model and the reference model, an intervention method distinct from reward model-based reinforcement learning optimization [11]. Related studies indicate that the way the model's probability distribution is adjusted is a crucial factor affecting capability retention after alignment [3, 4].

### 4.2. Data dependence and coverage limitations

The effectiveness of preference alignment is strongly influenced by the quality and coverage of training data. Although offline optimization methods represented by DPO alleviate the computational burden of reinforcement learning, their optimization remains constrained by pre-constructed static preference datasets. Thus, the range of behaviors the model can learn is limited by the data distribution. During actual training, as model parameters update, the policy model may produce outputs that deviate from the original data distribution, and offline preference data cannot provide additional feedback for these newly generated behaviors, affecting the model's adaptive capabilities under long-tail inputs and uncovered scenarios.

In contrast, online reinforcement learning methods like PPO continuously sample model generation results during training and provide dynamic evaluation signals through a reward model, allowing the policy to explore and adjust within the generation space. This mechanism mitigates the limitations of insufficient static data coverage while introducing additional training complexity and computational overhead. Thus, when supported by large-scale, high-quality preference data, online feedback-based RLHF provides stronger policy exploration and generalization potential, whereas offline optimization methods achieve greater training efficiency under resource constraints.

DOI: 10.54254/2755-2721/2026.BA36530

## 5. Conclusion

This paper focuses on large language model preference alignment methods and analyzes the sequential decision-making formulation, three-stage training mechanism, and key limitations of RLHF, including data costs, training instability, and reward model errors. On this basis, it discusses direct preference optimization methods like DPO, IPO, and KTO, as well as improvement schemes targeting preference signal acquisition and optimization processes like ReMax, RAFT, and RLAIF. The analysis shows that offline optimization methods such as DPO improve training efficiency by eliminating reward model training and reinforcement learning sampling processes, but their optimization performance remains constrained by preference data distribution and coverage. In contrast, RLHF relies on online policy optimization to maintain exploration capabilities and retains strong alignment potential when supported by high-quality preference data. At the same time, alignment methods based on model feedback reduce manual annotation costs, but feedback reliability and bias propagation remain issues to be resolved. Therefore, current preference alignment methods still face trade-offs among training costs, feedback quality, and capability retention. Nevertheless, this paper mainly reviews alignment mechanisms from existing studies and reports, without experiments comparing methods across different models and tasks. Future research could examine how different optimization mechanisms affect general capabilities like reasoning and code generation, and explore hybrid alignment methods that combine online exploration with offline optimization stability to improve large language models across safety, efficiency, and capability retention.

## References

[1] Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*. https://arxiv.org/abs/2204.05862
[2] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.
[3] Ji, J., et al. (2023). AI alignment: A comprehensive survey. *arXiv*. https://arxiv.org/abs/2310.19852
[4] Wang, Y., et al. (2023). Aligning large language models with human preferences: Methods and advancements. *arXiv*. https://arxiv.org/abs/2308.05374
[5] Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*.
[6] Ziegler, D. M., et al. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
[7] Stiennon, N., et al. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems, 33*, 3008-3021.
[8] Schulman, J., et al. (2017). Proximal policy optimization algorithms. *arXiv*. https://arxiv.org/abs/1707.06347
[9] Zheng, C., et al. (2023). Secrets of RLHF in large language models Part I: PPO. *arXiv*. https://arxiv.org/abs/2307.04964
[10] Casper, S., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv*. https://arxiv.org/abs/2307.15217
[11] Rafailov, R., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems, 36*.
[12] Azar, M. G., et al. (2023). A general theoretical framework for robust direct preference optimization. *arXiv*. https://arxiv.org/abs/2310.12036
[13] Ethayarajh, K., et al. (2024). KTO: Model alignment as prospect theoretic optimization. *arXiv preprint*. arXiv:2402.01306.
[14] Li, Z., et al. (2023). ReMax: A simple, effective, and efficient alternative to PPO for language model alignment. *arXiv*. https://arxiv.org/abs/2310.10505

DOI: 10.54254/2755-2721/2026.BA36530

[15] Dong, H., et al. (2023). RAFT: Rewarded action filtered fine-tuning for aligning large language models. *arXiv*. https://arxiv.org/abs/2304.06767
[16] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*. https://arxiv.org/abs/2212.08073
[17] Lee, H., et al. (2023). RLAIF: Scaling reinforcement learning from AI feedback with AI feedback. *arXiv*. https://arxiv.org/abs/2309.00267
[18] Yuan, W., et al. (2024). Self-reward language models. *arXiv*. https://arxiv.org/abs/2401.10020
[19] Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv*. https://arxiv.org/abs/2307.09288
[20] Team, G., Mesnard, T., Hardin, C., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv* preprint arXiv:2403.08295.