Adaptive-β Direct Preference Optimization for Stable Alignment of Large Language Models

Aref Ganjaee Sari
aref.ganjaee1@yahoo.com

ABSTRACT

Aligning large language models with human preferences is a central challenge in building reliable and trustworthy AI systems. Recently, Direct Preference Optimization (DPO) has been introduced as a simpler and more stable alternative to reinforcement learning from human feedback. Despite its structural advantages, DPO is highly sensitive to the choice of the preference scaling coefficient β, and using a fixed β can lead to training instability, excessive gradient growth, and undesirable drift from the reference policy. In this work, we propose Adaptive-β DPO, a novel approach in which the value of β is adjusted dynamically on a per-step basis using intrinsic training signals. These signals include the log-probability gap between preferred and non-preferred responses as an indicator of preference uncertainty, as well as a surrogate measure of Kullback–Leibler divergence to control deviation from the reference policy. The proposed mechanism increases β for hard preference instances while reducing it when divergence grows, thereby balancing effective preference learning with training stability. Experiments conducted on the GPT-2 model using the Helpful–Harmless dataset demonstrate that, under standard DPO with a fixed β, increasing β from 0.05 to 0.30 leads to a higher final loss value, increasingly negative reward margins, and more than a sixfold increase in gradient norms, while reward accuracy remains nearly unchanged. This behavior highlights the strong sensitivity of DPO to manual β tuning. In contrast, Adaptive-β DPO improves training stability and prevents excessive preference pressure without modifying the original DPO loss formulation. These results indicate that adaptive β tuning can significantly reduce reliance on manual hyperparameter selection and provide a more controllable and interpretable framework for preference alignment in language models.

KEYWORDS

Language Model Alignment; Direct Preference Optimization; Adaptive β Tuning; Preference Learning; Training Stability; KL Divergence

INTRODUCTION

Large language models (LLMs) have recently become a central component of many advanced artificial intelligence systems, demonstrating remarkable performance in tasks such as dialogue, reasoning, code generation, and text analysis. Despite these successes, LLMs are primarily trained on large-scale, heterogeneous human-generated data, which inherently contain a mixture of desirable behaviors, human errors, biases, and undesirable responses. As a result, aligning the behavior of language models with human preferences, values, and expectations has emerged as a fundamental challenge in the development of reliable LLM-based systems. The first widely successful framework for addressing this challenge was reinforcement learning from human feedback (RLHF), which demonstrated that human preferences can be effectively leveraged to guide model behavior [1]. However, RLHF suffers from several practical limitations, including high computational cost, the requirement to train a separate reward model, and the intrinsic instability of reinforcement learning algorithms. These drawbacks motivated the development of Direct Preference Optimization (DPO), which reformulates the RLHF objective to eliminate the need for both an explicit reward model and a reinforcement learning loop, enabling preference alignment to be achieved directly through a preference-based loss function [2]. Subsequent studies have shown that while DPO offers a simpler and more efficient alignment framework, the implicit rewards induced by its objective are not necessarily calibrated with respect to the base reward scale. To address this issue, Cal-DPO was introduced to calibrate these implicit rewards, improving learning stability and increasing the likelihood of preferred responses without adding significant complexity [3]. In parallel, safety has emerged as a critical concern in LLM alignment. Safe-DPO demonstrated that harmful behaviors can be mitigated by reordering preferences according to safety criteria, without relying on reinforcement learning or separate cost models [4]. Further efforts have focused on simplifying and generalizing the DPO framework. ORPO removed the reference policy to further streamline alignment [5], while Unified Preference Optimization showed that many preference-based methods can be analyzed as special cases of a unified theoretical framework [6]. Data efficiency has also received significant attention: Pre-DPO improved the utilization of preference data through informed use of the reference model [7], and SGDPO introduced self-guided learning to reduce reliance on human feedback [8]. Other works have explored temporal stability and inference efficiency. DiffPO, inspired by diffusion models, accelerated and stabilized alignment at inference time [9]. From a theoretical perspective, KTO incorporated insights from prospect theory to explicitly model human risk sensitivity and utility in preference alignment [10]. Distributional approaches such as D-RPO improved robustness under noisy preference data [11], while MoE-DPO leveraged mixture-of-experts architectures to better capture complex preference patterns [12]. Additional extensions, including Group Preference Optimization [13] and Soft Preference Optimization [14], addressed multi-objective alignment and soft preference scenarios. In specialized application domains, PLUM demonstrated that combining human preferences with actual code execution can significantly improve the alignment of code-oriented language models [15], and MDPO extended the DPO framework to multimodal models [16]. Other studies, such as Smaug [17] and related analyses of efficient alignment optimization [18], investigated failure modes and robustness issues in preference-based methods. More recent work has emphasized diverse, multi-objective, and adaptive preferences. Auto-Weighted Group RPO [19], Contrastive Preference Optimization [20], and Self-Rewarding Language Models [21] each address limitations of traditional alignment from different perspectives. Investigations into the role of the reference policy in DPO [22], comprehensive surveys [23-26], accelerated alignment methods [27], diverse preference

modeling [28], and reward-aware frameworks [29] further illustrate that preference-based alignment remains an active and rapidly evolving research area. Adaptive strategies in multimodal settings [30] and representation engineering [31] have likewise highlighted the importance of dynamically regulating model behavior. Despite these extensive advances, a fundamental limitation is shared across nearly all preference-based alignment methods: the strength of preference enforcement is controlled by a fixed scalar coefficient β. This parameter plays a critical role in determining gradient magnitudes, convergence stability, and the degree of Kullback–Leibler (KL) divergence from the reference policy. Relying on a constant β across all training stages and data instances ignores the inherently dynamic nature of preference uncertainty and training dynamics. To address this limitation, we introduce Adaptive-β DPO, a method that dynamically adjusts β based on intrinsic training signals and optimization conditions. By enabling adaptive control over preference strength, the proposed approach aims to simultaneously improve training stability and alignment effectiveness, without altering the original DPO loss structure.

RELATED WORK

With the rapid advancement of large language models (LLMs) in recent years, numerous methods have been proposed to align these models with human preferences. Table 1 provides a structured overview of the most influential preference-based alignment approaches. Before presenting the table, we briefly review the evolutionary trajectory of these methods. Reinforcement Learning from Human Feedback (RLHF) was the first widely adopted solution for aligning language models with human values and preferences [1]. Despite its success, RLHF suffers from several practical limitations, including the complexity of reward model construction, high computational cost, and the extensive requirement for human-labeled data, which restrict its scalability to modern large-scale models. To address these limitations, Rafailov et al. introduced Direct Preference Optimization (DPO) [2], which reformulates preference learning directly at the policy level and eliminates the need for an explicit reward model. By reducing alignment to a cross-entropy–based loss function, DPO marked a turning point in the literature and stimulated a broad line of subsequent research. Following the introduction of DPO, the calibration of implicit rewards emerged as a key challenge. Cal-DPO demonstrated that the implicit rewards derived from DPO may be misaligned with the base reward scale, potentially reducing the probability of preferred responses and destabilizing model behavior. By introducing a simple corrective mechanism, Cal-DPO improved output stability and alignment performance without significantly increasing complexity [3]. Several studies focused on strengthening the theoretical foundations of DPO. ORPO simplified the alignment process by removing the reference policy π_{ref} [5], while KTO incorporated insights from prospect theory to explicitly model human risk sensitivity and utility within preference optimization [10]. Beyond theoretical advances, some works targeted improved performance under limited data. Group Preference Optimization emphasized group-level and multi-objective preference scenarios [13], whereas Soft Preference Optimization mitigated preference conflicts by smoothing the preference distribution [14]. In parallel, Pre-DPO significantly improved preference data efficiency by leveraging a guiding reference model [7]. Preference-based alignment methods have also been extended to specialized application domains. PLUM demonstrated that combining human preferences with real code execution substantially improves alignment in code-oriented language models [15]. Similarly, MDPO generalized the DPO framework to multimodal models, showing its applicability to vision–language tasks [16]. More recent research has emphasized safety, stability, and robustness. DiffPO, inspired by diffusion models, improved both alignment speed and output stability [9]. SGDPO introduced self-guided learning, demonstrating that part of the preference signal can be generated autonomously from model behavior, thereby reducing reliance on human feedback [8]. From a safety perspective, Safe-DPO showed that reordering preferences according to safety criteria can produce safer outputs without reward models or reinforcement learning loops [4]. Smaug addressed issues such as over-penalization and training collapse by stabilizing gradient behavior during preference optimization [17]. Additionally, D-RPO adopted a distributional perspective to improve robustness under noisy and imbalanced preference data [11]. From a unifying theoretical standpoint, Unified Preference Optimization demonstrated that many preference-based alignment methods can be interpreted as variants of a shared objective function [6]. Other work focused on improving the temporal and memory efficiency of preference alignment [18]. Finally, Mix/MoE-DPO leveraged mixture-of-experts architectures to better model complex, high-dimensional preference patterns and improve performance in multi-faceted alignment scenarios [12].

Table 1. Comparative overview of preference-based alignment methods and remaining challenges

No.	Method	Year	Publication	Model / Framework	Algorithm	Key Results	Short Summary	Main Limitation
1	DPO	2023	NeurIPS	GPT-J, Pythia	Cross-Entropy DPO	10–12% win-rate improvement on summarization and dialogue	Introduces DPO as a simple RLHF alternative without an explicit reward model	Fixed β leads to instability, high KL divergence, and policy drift

2	Cal-DPO	2024	NeurIPS	GPT-J, Helpful–Harmless	Calibrated DPO	↓25% KL divergence, ↑8% training stability	Calibrates implicit rewards to prevent degradation of preferred response probability	β remains fixed; limited robustness under noisy preferences
3	Safe-DPO	2025	arXiv	Reddit + Helpful–Harmless	Safety-aware DPO	↓23% reward error under noisy data	Reorders preferences using safety criteria without reward models	No adaptive β; safety improved but preference strength remains static
4	Pre-DPO	2025	arXiv	Anthropic-HH	Guided DPO	↑15% data efficiency, ↓5% loss	Improves data utilization via a guiding reference model	Fixed β; dynamic training behavior not considered
5	SGDPO	2025	ACL	TL;DR, Helpful–Harmless	Self-Guided DPO	↑9% win-rate, ↑12% stability	Reduces reliance on human feedback through self-guided learning	β is static; limited control over training dynamics
6	DiffPO	2025	ACL	TL;DR	Diffusion-based DPO	↓20% inference time with comparable output quality	Accelerates and stabilizes alignment using diffusion-inspired structures	β remains fixed; potential KL drift
7	ORPO	2024	arXiv	LLaMA	ORPO	Reduced dependence on reference model	Removes the reference policy to simplify alignment	No dynamic β; KL divergence not explicitly controlled
8	KTO	2024	arXiv	Helpful–Harmless	Prospect-Theoretic Optimization	Improved sensitivity to human risk and utility	Integrates prospect theory into preference optimization	Fixed β; KL control remains implicit
9	D-RPO	2024	arXiv	Anthropic-HH	Distributionally Robust DPO	↑12% robustness score	Improves stability under noisy and imbalanced preference data	β is not adaptive despite improved robustness
10	Mix/MoE-DPO	2025	arXiv	Helpful–Harmless	Mixture-of-Experts DPO	↑7% accuracy, ↓10% overfitting	Models complex preferences via MoE architectures	Fixed β causes gradient instability in complex settings

A comparative analysis of the methods summarized in Table 1 reveals that while Direct Preference Optimization (DPO) represents a major step toward simplifying language model alignment, training stability remains a fundamental challenge across most of its variants and subsequent extensions. Nearly all preference-based methods—including DPO, Cal-DPO, Safe-DPO, ORPO, KTO, D-RPO, and more advanced approaches such as MoE-DPO—rely on a fixed scalar coefficient β to control the strength of preference enforcement and the degree of deviation from the reference policy. Reported empirical results across these studies consistently indicate that the choice of β has a direct and significant impact on gradient norms, convergence behavior, the magnitude and direction of KL divergence, and ultimately overall training stability. While several works attempt to mitigate practical limitations of DPO through objective reformulation (e.g., KTO), preference reordering (e.g., Safe-DPO), or improved data efficiency (e.g., Pre-DPO), none directly address the problem of dynamic, step-by-step adjustment of β. In most existing methods, β is selected prior to training and remains unchanged throughout the entire optimization process. This static treatment of β is fundamentally misaligned with the inherently dynamic nature of preference learning, which involves a mixture of easy and hard samples, unstable early training phases, and more stable later stages. Existing empirical findings implicitly suggest that a large β can induce excessive preference pressure, rapid growth of gradient norms, and undesirable increases in KL divergence, whereas a small β may weaken preference learning and lead to slow or insufficient convergence. Despite these observations, the current literature lacks a data-driven and self-adaptive mechanism for tuning β that simultaneously accounts for preference uncertainty and deviation from the reference policy. This gap indicates that improving training stability and effectively controlling KL divergence in preference-based alignment requires treating β not as a fixed hyperparameter, but as a dynamic variable conditioned on the training state. Motivated by this insight, we propose Adaptive-β DPO, in which the value of β is adjusted adaptively based on intrinsic training signals. This design enables stronger learning on difficult preference instances while adopting a more conservative behavior under unstable conditions, thereby balancing effective preference learning with robust training stability.

METHOD: ADAPTIVE-B DIRECT PREFERENCE OPTIMIZATION

Problem Setting

Direct Preference Optimization (DPO) is an effective framework for aligning language models with human preferences without training a separate reward model. In DPO, the strength of preference enforcement is controlled by a scalar coefficient β. In the standard formulation, β is fixed throughout training. While this design simplifies optimization, prior studies and empirical observations show that the choice of β plays a critical role in training stability. Large values of β may induce excessive preference pressure, sharp growth of gradient norms, and increased divergence from the reference policy, whereas small values of β can weaken preference learning and slow convergence. This sensitivity makes DPO heavily dependent on manual tuning of a critical hyperparameter. To address this limitation, we propose Adaptive-β DPO, which introduces a dynamic mechanism for adjusting β during training. The key idea is that the strength of preference enforcement should adapt to the instantaneous training state of the model rather than remain constant.

Preference Modeling and DPO Objective

We assume a preference dataset consisting of triplets (x, y_w, y_l), where x is the input prompt, y_W is the preferred response, and y_l is the non-preferred response. Let π_θ denote the current policy and π_{ref} the reference policy. For each sample, we define the log-probability difference under the current policy as:

$$\Delta_\theta(\mathrm{x}, y_w, y_l) = \log\, \pi_\theta\,(y_w|x) - \log\, \pi_\theta\,(y_l|x) \tag{1}$$

This quantity measures how confidently the model distinguishes preferred from non-preferred responses. Large positive values indicate high confidence in the human preference, while small or negative values reflect uncertainty and increased sample difficulty. We therefore use Δ_θ as an intrinsic signal for preference uncertainty. The standard DPO loss is defined as:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}[\log\, \sigma\left(\beta(\Delta_\theta - \Delta_{ref})\right)] \tag{2}$$

where $\sigma(\cdot)$is the sigmoid function and β controls the scale of preference enforcement. This coefficient directly affects gradient magnitudes and convergence stability.

Controlling Deviation from the Reference Policy

To prevent excessive drift from the reference policy, we introduce a surrogate measure of the Kullback–Leibler divergence between the current and reference policies:

$$\widehat{D}_{KL} = D_{KL}\left(\pi_\theta - \pi_{ref}\right) \tag{3}$$

This term acts as a control signal that reflects how far the model deviates from its initial behavior during training and helps mitigate numerical instability and policy drift.

Adaptive β Update Rule

In Adaptive-β DPO, the coefficient β is updated dynamically using a combination of the preference uncertainty signal and the divergence from the reference policy:

$$\beta_{t+1} = clip(\beta_t + \eta_\beta g(\Delta_\theta) - \gamma.h\left(\widehat{D}_{KL}\right), \beta_{min}, \beta_{max} \tag{4}$$

Here, $g(\cdot)$ is an increasing function of preference uncertainty, encouraging stronger learning on difficult samples, and $h(\cdot)$ is an increasing function of divergence from the reference policy, reducing β when excessive deviation occurs. The interval $[\beta_{min}, \beta_{max}]$ ensures safe operation and prevents extreme behaviors. Intuitively, when the model exhibits low confidence in preference discrimination, β is increased to strengthen learning pressure. Conversely, when divergence from the reference policy grows, β is reduced to preserve training stability. The updated β is directly used in the DPO loss (Eq. 2), enabling adaptive control over gradient magnitudes without altering the original loss structure.

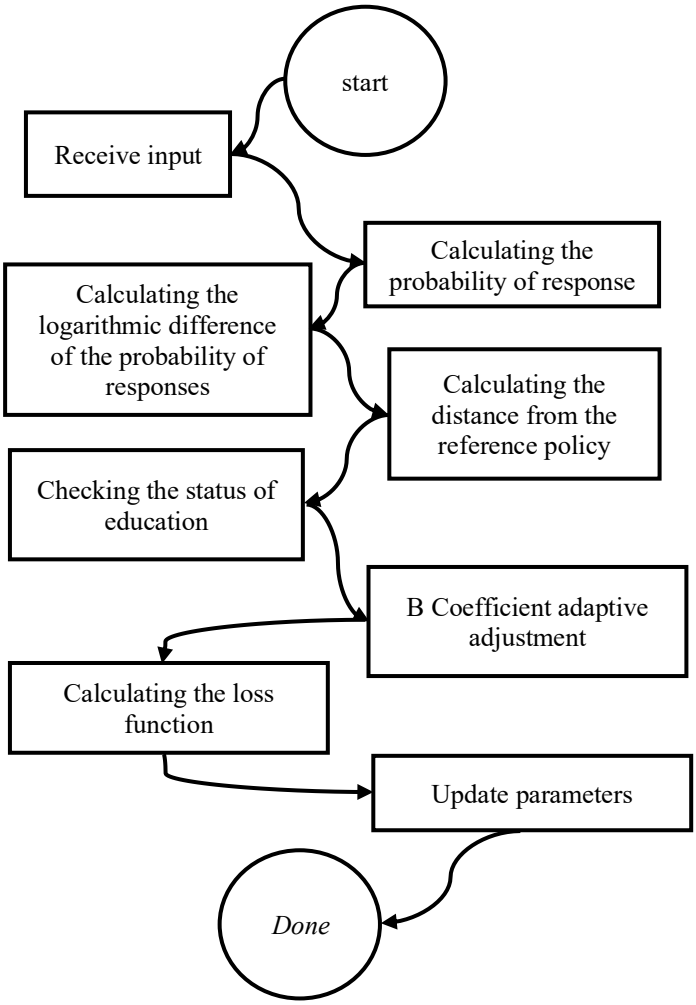


FIGURE 1. ILLUSTRATES THE OVERALL WORKFLOW OF THE PROPOSED ADAPTIVE-B DPO ALGORITHM

The process begins by receiving preference data and computing response probabilities under the current policy. The log-probability difference between preferred and non-preferred responses is then calculated to assess preference uncertainty. Simultaneously, the divergence from the reference policy is evaluated as a stability control signal. Based on these two signals, the coefficient β is adjusted adaptively and used to compute the DPO loss. Finally, model parameters are updated, and this process is repeated until convergence.

EVALUATION AND RESULTS

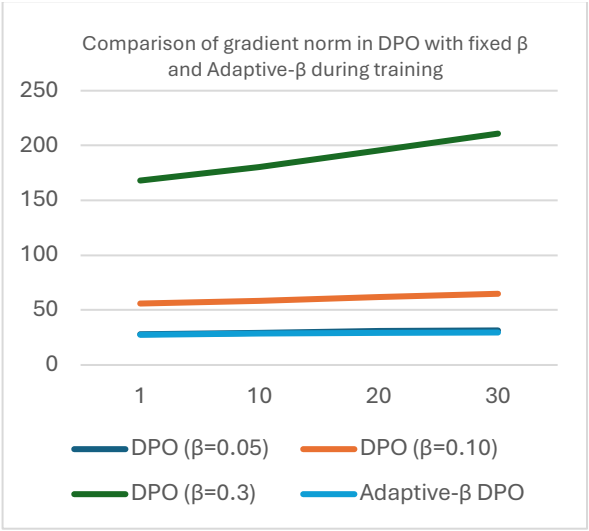
In this section, we evaluate the performance of the proposed Adaptive-β DPO method in comparison with the standard Direct Preference Optimization (DPO) using a fixed β coefficient. The primary objective of this evaluation is not to compare final response quality, but rather to analyze training stability, gradient behavior, and control over deviation from the reference policy under controlled conditions. We conduct experiments on the widely used Helpful–Harmless (HH) preference dataset, which consists of triplets (x, y_w, y_l) including an input prompt, a preferred response, and a non-preferred response. This dataset is commonly adopted in DPO-based alignment studies. Across all experiments, the data structure, preprocessing pipeline, base model architecture (GPT-2), and training configurations are kept identical to ensure a fair and reproducible comparison. To assess performance, we employ a set of complementary metrics focusing on numerical stability and preference-learning dynamics. First, we analyze the training loss and its evolution over optimization steps. Second, reward-based metrics—including reward margin, reward chosen, and reward rejected—are examined to quantify the model's ability to distinguish preferred from non-preferred responses. In addition, the gradient norm is monitored as a key indicator of numerical stability, since excessive growth in gradient magnitude often signals over-enforcement of preferences and a heightened risk of optimization instability. For the proposed Adaptive-β DPO, the dynamic behavior of β throughout training is also recorded and analyzed to investigate the relationship between model uncertainty and preference enforcement strength. To study the sensitivity of standard DPO to the choice of β, we run the fixed-β variant with three values: β = 0.05, β = 0.10, and β = 0.30.

Table 2. Comparison of standard DPO with different fixed β values

β value	Final Loss	Final Reward Margin	Final Gradient Norm
0.05	0.7373	−0.0851	31.52
0.10	0.7792	−0.1605	64.87
0.30	0.9231	−0.3901	210.94

Table 2 summarizes the experimental results, reporting the final training loss, reward margin, and gradient norm at the end of training. The results show that increasing β leads to a higher final loss and drives the reward margin toward more negative values. Moreover, the gradient norm grows rapidly as β increases, reaching extremely large values when β = 0.30. This behavior indicates that increasing β does not necessarily improve preference learning and can instead impose excessive preference pressure, resulting in numerical instability during training. In contrast, reward accuracy remains approximately constant across all β values, highlighting that manual tuning of β constitutes a sensitive and potentially unstable design choice in standard DPO.

Figure 2. Comparison of DPO with fixed and adaptive β values



Analysis of Gradient Norm Dynamics

Figure 2 compares the evolution of the gradient norm during training for standard DPO with fixed β values and the proposed Adaptive-β DPO. As shown, fixed-β DPO exhibits a clear sensitivity to the choice of β. When β is small (β = 0.05), gradient magnitudes remain relatively low but grow steadily, indicating weak preference pressure and limited learning strength. Increasing β to 0.10 results in moderately larger gradients, while a further increase to β = 0.30 causes a substantial escalation in gradient norms throughout training, reflecting excessive preference enforcement and a heightened risk of numerical instability. In contrast, Adaptive-β DPO maintains consistently low and stable gradient norms across training steps. Unlike fixed-β variants, which apply uniform preference pressure regardless of training state, the adaptive mechanism dynamically regulates β based on model uncertainty and deviation from the reference policy. This adaptive behavior prevents uncontrolled gradient growth while preserving effective preference learning. The observed stability demonstrates that Adaptive-β DPO successfully decouples training stability from manual β selection and mitigates the over-optimization effects seen in large fixed-β settings.

Table 3. Comparison of Adaptive-β DPO with fixed-β preference optimization methods

Method	β Setting	KL Control	Gradient Stability	Reward Model Required	Primary Focus
DPO [2]	Fixed	Indirect	Unstable for large β	No	Simplicity as an RLHF alternative
Cal-DPO [3]	Fixed	Partial	Moderate	No	Implicit reward calibration
Safe-DPO [4]	Fixed	Partial	Moderate	No	Output safety
D-RPO [11]	Fixed	Partial	Improved	No	Robustness to noisy data
Adaptive-β DPO	Adaptive	Active	Stable	No	Training stability + KL control

As shown in Table 3, most prior preference-based alignment methods rely on a fixed β coefficient, while focusing on complementary aspects such as reward calibration, safety, or robustness to noise. In contrast, Adaptive-β DPO is the first framework that performs dynamic, data-driven adjustment of β, explicitly targeting both gradient stability and control of KL divergence from the reference policy—without requiring a separate reward model or reinforcement learning.

In Adaptive-β DPO, the value of β evolves dynamically throughout training based on intrinsic model signals. Analysis of training logs shows that β typically starts from relatively small values during early training stages and gradually increases as learning progresses, approaching a predefined upper bound. This behavior reflects conservative preference enforcement during unstable early phases and progressively stronger preference pressure during later, more stable stages of training. Further inspection of the log-probability difference between preferred and non-preferred responses reveals substantial fluctuations, including negative values at certain steps, indicating the presence of high-uncertainty preference samples. Under such conditions, the adaptive mechanism modulates preference strength in a controlled manner, preventing the application of uniform and potentially destabilizing preference pressure. At the same time, the surrogate KL divergence remains within a low range throughout training, demonstrating effective control over deviation from the reference policy. Overall, the experimental results indicate that Adaptive-β DPO achieves a more stable and controllable training behavior than fixed-β DPO, without altering the original DPO loss structure. Importantly, the goal of these experiments is not to evaluate final response quality, but to analyze training stability and the dynamic behavior of preference enforcement under controlled conditions. The findings suggest that adaptive regulation of β can substantially improve training stability without increasing computational complexity. It should be noted that the current evaluation is conducted at a limited scale, primarily to validate the implementation and analyze the dynamic behavior of β. Future studies using larger datasets, advanced qualitative metrics such as win-rate, and more precise measurements of KL divergence could provide a more comprehensive assessment of the proposed approach.

CONCLUSION

In this work, we investigated the sensitivity of Direct Preference Optimization (DPO) to the choice of the preference strength coefficient β and proposed an adaptive mechanism for regulating this parameter during training. Empirical analysis shows that using a fixed β—particularly at larger values—can lead to rapid growth in gradient norms, increased divergence from the reference policy, and reduced training stability, without delivering meaningful improvements in preference learning. In contrast, the proposed Adaptive-β DPO method dynamically and data-dependently adjusts β according to the instantaneous training state of the model. Experimental observations demonstrate that this approach maintains gradient norms within a stable range, avoids the severe oscillations observed in fixed-β DPO, and enables more controlled deviation from the reference policy. As a result, the training process exhibits more predictable and interpretable behavior. A key advantage of Adaptive-β DPO is that it improves training stability without modifying the original DPO loss function, without introducing a separate reward model, and without relying on reinforcement learning. From this perspective, Adaptive-β DPO can be viewed as a simple yet effective extension of standard DPO that reduces dependence on manual tuning of a critical hyperparameter. These findings suggest that adaptive regulation of preference strength plays an important role in stabilizing preference-based alignment algorithms and provides a promising foundation for developing more robust variants of DPO in future work.

References

1. Long, O., et al., *Training language models to follow instructions with human feedback.* Advances in neural information processing systems, 2022. **35**: p. 27730–27744.
2. Rafailov, R., et al., *Direct preference optimization: Your language model is secretly a reward model.* Advances in neural information processing systems, 2023. **36**: p. 53728–53741.
3. Xiao, T., et al., *Cal-dpo: Calibrated direct preference optimization for language model alignment.* Advances in Neural Information Processing Systems, 2024. **37**: p. 114289–114320.
4. Kim, G.-H., et al., *SafeDPO: A simple approach to direct preference optimization with enhanced safety.* arXiv preprint arXiv:2505.20065, 2025.
5. Hong, J., N. Lee, and J. Thorne, *Orpo: Monolithic preference optimization without reference model.* arXiv preprint arXiv:2403.07691, 2024.
6. Badrinath, A., P. Agarwal, and J. Xu, *Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier.* arXiv preprint arXiv:2405.17956, 2024.
7. Pan, J., et al., *Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model.* arXiv preprint arXiv:2504.15843, 2025.
8. Zhu, W., et al., *SGDPO: Self-Guided Direct Preference Optimization for Language Model Alignment.* arXiv preprint arXiv:2505.12435, 2025.
9. Chen, R., et al. *DiffPO: Diffusion-styled Preference Optimization for Inference Time Alignment of Large Language Models*. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025.
10. Ethayarajh, K., et al., *Kto: Model alignment as prospect theoretic optimization, 2024.* URL https://arxiv. org/abs/2402.01306.
11. Wu, J., et al., *Towards robust alignment of language models: Distributionally robustifying direct preference optimization.* arXiv preprint arXiv:2407.07880, 2024.
12. Bohne, J., et al., *Mix-and MoE-DPO: A Variational Inference Approach to Direct Preference Optimization.* arXiv preprint arXiv:2510.08256, 2025.
13. Zhao, S., J. Dang, and A. Grover, *Group preference optimization: Few-shot alignment of large language models.* arXiv preprint arXiv:2310.11523, 2023.
14. Sharifnassab, A., et al., *Soft preference optimization: Aligning language models to expert distributions.* arXiv preprint arXiv:2405.00747, 2024.
15. Zhang, D., et al., *$\textbf{PLUM} $: Improving Code LMs with Execution-Guided On-Policy Preference Learning Driven By Synthetic Test Cases.* arXiv preprint arXiv:2406.06887, 2024.
16. Wang, F., et al., *mdpo: Conditional preference optimization for multimodal large language models.* arXiv preprint arXiv:2406.11839, 2024.
17. Pal, A., et al., *Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024.* URL https://arxiv. org/abs/2402.13228.
18. Ji, H., *Towards efficient exact optimization of language model alignment (2024).* URL https://arxiv. org/abs/2402.00856. **2402**.
19. Ichihara, Y. and Y. Jinnai. *Auto-Weighted Group Relative Preference Optimization for Multi-Objective Text Generation Tasks*. in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2025.
20. Xu, H., et al., *Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.* arXiv preprint arXiv:2401.08417, 2024.
21. Yuan, W., et al. *Self-rewarding language models*. in *Forty-first International Conference on Machine Learning*. 2024.
22. Liu, Y., P. Liu, and A. Cohan, *Understanding reference policies in direct preference optimization.* arXiv preprint arXiv:2407.13709, 2024.
23. Xiao, W., et al., *A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications.* arXiv preprint arXiv:2410.15595, 2024.
24. Winata, G.I., et al., *Preference tuning with human feedback on language, speech, and vision tasks: A survey.* Journal of Artificial Intelligence Research, 2025. **82**: p. 2595–2661.
25. Liang, X., et al., *ROPO: Robust Preference Optimization for Large Language Models.* arXiv preprint arXiv:2404.04102, 2024.
26. Liu, S., et al., *A survey of direct preference optimization.* arXiv preprint arXiv:2503.11701, 2025.
27. He, J., H. Yuan, and Q. Gu, *Accelerated preference optimization for large language model alignment.* arXiv preprint arXiv:2410.06293, 2024.
28. Zeng, D., et al. *On diversified preferences of large language model alignment*. in *Findings of the association for computational linguistics: EMNLP 2024*. 2024.
29. Sun, S., et al., *Reward-aware preference optimization: A unified mathematical framework for model alignment.* arXiv preprint arXiv:2502.00203, 2025.
30. Lu, J., et al., *Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization.* arXiv preprint arXiv:2504.15619, 2025.
31. Liu, W., et al. *Aligning large language models with human preferences through representation engineering*. in *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.