CoPO: Contrastive Preference Optimization via On-Policy Reward Trajectory Alignment

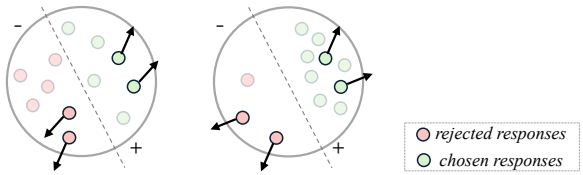
Anonymous ACL submission

Abstract

Preference optimization has become a standard paradigm for aligning large language models (LLMs) with human preferences. Existing finegrained preference optimization methods usually improve preference signal utilization beyond sequence-level objectives by introducing token-aware or trajectory-level supervision. However, existing methods optimize preference margins over observed responses, while autoregressive generation depends on decoding trajectories. This optimization mismatch causes supervision gradually narrows effective preference regions and leads to preference collapse. To address this issue, we propose Contrastive Preference Optimization (CoPO), a preference optimization framework that aligns preference supervision with generation behavior through reward trajectory alignment. Specifically, CoPO introduces auxiliary anchor responses sampled from the current policy and contrastively aligns their token-level implicit reward trajectories toward preferred responses while separating them from rejected ones. Our method expands the coverage of preference-consistent reward regions. Experiments on seven benchmarks demonstrate that CoPO consistently improves preference alignment across different LLM backbones and multi-backbone preference data.

1 Introduction

Aligning large language models (LLMs) with human preferences has become a central objective in post-training. Reinforcement learning from human feedback (RLHF) has achieved strong empirical performance by learning reward models and optimizing policies through reinforcement learning (Zheng et al., 2025; Gao et al., 2025; Xu and Ding, 2026; Qi et al., 2026). Although effective, these approaches typically require additional reward modeling and iterative policy optimization, resulting in increased training complexity and engineering overhead.



(a) $\mathcal{L}_{DPO}$ (b) $\mathcal{L}_{CoPO}$ (ours)

Figure 1: Illustration of preference supervision under DPO and CoPO. CoPO introduces policy-generated responses as anchors and aligns the model's generated trajectories toward preferred responses while separating them from rejected ones. As training progresses, the model output distribution gradually shifts toward the preferred region, leading to more preference-consistent generation.

To simplify this, preference optimization methods (Rafailov et al., 2023; Meng et al., 2024) optimize preference objectives directly through implicit rewards. Recent studies further improve preference optimization by introducing fine-grained preference supervision beyond sequence-level objectives. Token-aware methods (Zeng et al., 2024; Zhu et al., 2025; Liu et al., 2025) refine preference signals through token-level reward estimation, while trajectory-based methods (Yang et al., 2026) incorporate structural constraints over generated responses and intermediate reward behaviors. These approaches improve preference signal utilization and provide richer supervision during optimization.

However, existing preference optimization methods suffer from a fundamental limitation: optimization mismatch between preference objectives and autoregressive generation objectives. Existing methods optimize preference learning by enlarging the relative margin between chosen and rejected responses. Although this improves preference scores on observed response pairs, it does not explicitly constrain the model's actual generation trajectory. During inference, language models generate responses autoregressively by maximizing

token probabilities step by step, making generation quality depend on the entire token trajectory rather than a few preference comparisons. As illustrated in Figure 1(a), optimizing preference margins alone mainly pushes selected responses farther apart in preference space, while the generated distribution may not consistently move toward the preferred region. As training progresses, optimization gradually concentrates on a limited subset of high-reward tokens and weakens constraints over the remaining generated tokens, causing preference collapse.

To achieve this, we introduce policy-generated signals into preference optimization, enabling more stable and preference-consistent generation while preserving the simplicity and efficiency of direct preference optimization. Specifically, as shown in Figure 1(b), CoPO introduces model-generated responses as implicit anchors and represents them as token-level reward trajectories. Instead of enlarging preference margins on static response pairs, CoPO contrastively pulls generated trajectories toward preferred responses while separating them from rejected ones. This trajectory-level supervision transfers preference optimization from response comparison to generation behavior itself, alleviating optimization mismatch and encouraging more preference-consistent generation. To further stabilize training, we introduce a quality-aware calibration strategy. Importantly, CoPO preserves the simplicity and efficiency of direct preference optimization without requiring additional reward models or reinforcement learning.

We propose a new Contrastive Preference Optimization (CoPO) framework to align preference supervision directly with generation behavior through reward trajectory alignment. Specifically, CoPO introduces auxiliary anchor responses generated from the current policy and represents them as token-level implicit reward trajectories. Instead of optimizing absolute preference scores, CoPO contrastively aligns anchor trajectories toward preferred responses while separating them from rejected responses, transferring supervision from response-level comparison to generation behavior itself. To further stabilize optimization, we introduce a quality-aware calibration strategy that adaptively adjusts supervision according to response quality. Importantly, CoPO preserves the simplicity and efficiency of direct preference optimization without requiring additional reward models or reinforcement learning optimization.

We conduct experiments on seven benchmarks across reasoning and knowledge. Experimental results show that CoPO consistently improves preference alignment across different LLM backbones. Further analysis demonstrates that our method exhibits strong generalization on multi-backbone preference data.

Our contributions are summarized as follows: 1) We identify a key limitation of existing fine-grained preference optimization methods: preference supervision lacks explicit constraints over the model's current generation trajectory, which may lead to preference collapse. To address this, we introduce a new perspective of regulating generation behavior during preference optimization via policy-generated anchor signals. 2) We propose Contrastive Preference Optimization (CoPO) to align preference supervision with evolving policy behavior, without requiring reward models or RL optimization. 3) Experiments on seven benchmarks demonstrate that CoPO consistently improves preference alignment. Our method can generalize well across model scales and preference data generated from different LLM backbones.

2 Preliminaries

Problem Formulation Preference optimization aims to align a policy model with human preferences by increasing the likelihood of preferred responses over dispreferred ones. Given a preference dataset $\mathcal{D} = \{(x, y^-, y^+)\}$, where x denotes the input prompt and y^+ and y^- denote preferred and dispreferred responses respectively, the objective is to learn a policy π_θ that assigns higher probability to preferred responses: $\pi_\theta(y^+|x) > \pi_\theta(y^-|x)$.

Direct Preference Optimization (DPO) Instead of training an explicit reward model, Direct Preference Optimization (DPO) (Rafailov et al., 2023) learns a policy directly from human preference data by reparameterizing the reward R. The training objective of DPO can be defined as: $\mathcal{L}_{\text{DPO}} = -\log\sigma\left(\beta\left(R(x, y^+) - R(x, y^-)\right)\right)$, where σ indicates the sigmoid function. π_θ is the policy model under training, and π_{ref} is the reference model. β is a hyperparameter that controls the deviation from the reference policy.

3 Method

3.1 Overview

We propose Contrastive Preference Optimization (CoPO), an offline preference optimization frame-

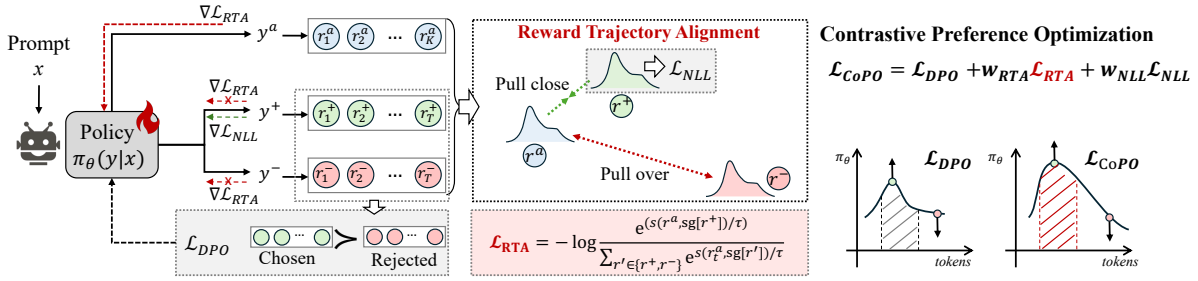


Figure 2: The overall of the proposed Contrastive Preference Optimization (CoPO). It aims to align preference supervision with evolving policy behavior, without requiring reward models or RL optimization. Specifically, CoPO introduces policy-generated anchor responses and performs reward trajectory alignment in implicit reward space. By aligning anchor trajectories toward preferred responses and separating them from rejected responses, CoPO expands the coverage of preference-consistent reward regions. As shown in the shaded area on the right, more policy-generated tokens are encouraged to fall into regions whose implicit rewards are closer to preferred trajectories.

work that reduces the mismatch between training-time preference supervision and generation-time behaviors. The overall architecture of CoPO is shown in Figure 2. Unlike existing DPO-based methods that optimize only observed preference responses, CoPO introduces sampled responses generated from the current policy as optimization anchors and performs reward trajectory alignment in token-level implicit reward space. To stabilize preference transfer, CoPO further introduces a quality-aware calibration strategy that adaptively emphasizes reliable preference pairs and selectively reinforces high-confidence preferred responses.

3.2 Normalized DPO

Given a human preference dataset $\mathcal{D} = \{(x, y^+, y^-)\}$, DPO defines implicit reward r through the likelihood ratio between the current policy π_θ and a frozen reference model π_{ref}. Since sequence-level rewards accumulate across generation length, DPO often exhibits undesired length sensitivity. To avoid this, we follow Meng et al. (2024) and normalize sequence rewards by response length:

$$R(x,y) = \frac{1}{|y|}\sum_{t=1}^{|y|} r(x, y_t) = \frac{1}{|y|}\sum_{t=1}^{|y|} \log \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}. \quad (1)$$

where $|y|$ is the length of response y. The normalized DPO objective can be refined as,

$$\mathcal{L}_{DPO} = -\log \sigma\left(\beta(R(x,y^+) - R(x,y^-))\right) \quad (2)$$

This operation mitigates undesired length effects and provides a more stable reward space for preference alignment.

3.3 Vanilla CoPO: Contrastive Preference Optimization

Existing token-aware methods mainly reweight preference losses or introduce margin constraints. Instead, CoPO directly optimizes relationships among generated reward trajectories. Specifically, CoPO first introduces implicit anchor rewards, allowing optimization to directly operate on responses generated from the current policy. Then, the soft-alignment loss optimizes the policy to align anchor trajectories adaptively.

Implicit Anchor Rewards Formally, define the reward trajectories in DPO are: $[r(x, y_1), ..., r(x, y_T)]$, where $r(x, y_t)$ is the token-level implicit reward. Reward trajectories characterize generation behavior in implicit reward space. We sample auxiliary anchor responses from the current policy, i.e.,

$$y^a_{1:K} \sim \pi_\theta(\cdot|x), \quad (3)$$

where K is the anchor length. The reward trajectories of anchor response y^a is $[r(x, y^a_1), ..., r(x, y^a_H)]$. Compared with static preference responses, anchor trajectories directly reflect the model's own sampled outputs and enable behavior-level preference optimization.

Reward Trajectory Alignment We design reward trajectory alignment (RTA) to align generated reward trajectories toward preferred responses

while separating them from rejected responses. The loss function of RTA is defined as:

$$\mathcal{L}_{\text{RTA}} = \mathbb{E}_{t\sim T} - \log \frac{\exp(s\left(r_t^a, \text{sg}[r_t^+]\right)/\tau)}{\sum_{r'_t \in \{r_t^+, r_t^-\}} \exp(s(r_t^a, \text{sg}[r'_t])/\tau)} \tag{4}$$

where $s(\cdot)$ a pairwise similarity function, i.e., cosine similarity, which can be seen as the dot product with L_2 normalization. $\text{sg}[\cdot]$ refers to the stop-gradient operator. This asymmetric optimization avoids representation collapse and ensures the stability of optimization. Unlike DPO, which only optimizes observed preference responses, the reward trajectory alignment allows preference supervision to propagate to policy-generated trajectories. As anchor responses gradually move toward preferred trajectories in implicit reward space, tokens that were previously weakly aligned or outside the observed preference pairs can become closer to preferred reward regions. Consequently, CoPO increases the coverage of preference-consistent trajectories during generation. This effect is illustrated as the enlarged shaded region in Figure 2.

3.4 Quality-aware CoPO

Preference pairs exhibit varying confidence and optimization value. Directly aligning all sampled behaviors may amplify noisy supervision and weaken behavior correction. Therefore, we apply the quality-aware calibration strategy to enable high-confidence preference pairs for more stable alignment.

First, we apply a margin-normalized weighting on preference pair data, i.e., $w_{\text{RTA}} = \lambda_{\text{RTA}} \times \frac{m_i - \mu}{\sigma + \epsilon}$ where m_i denotes the reward margin, μ and σ are batch-level statistics. High-confidence pairs contribute stronger alignment signals while noisy supervision is suppressed.

Besides, to strength preferred generations and improves response stability, we regularize preferred responses via the negative log-likelihood term:

$$\mathcal{L}_{NLL} = -\frac{1}{|y|} \sum_{t=1}^{|y|} \mathbf{1}(S(y_t^+) > \overline{S(y^+)}) \log \pi_\theta(y_t^+ | x, y_{<t}), \tag{5}$$

where $S(y_t^+)$ is the reward scoring by reward model in the dataset. This term emphasizes challenging and high-quality positive samples.

The final optimization objective combines normalized DPO alignment, reward trajectory alignment, and preference calibration, i.e.,

$$\mathcal{L}_{\text{CoPO}} = \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}_{\text{DPO}} + w_{\text{RTA}}\mathcal{L}_{\text{RTA}} + w_{\text{NLL}}\mathcal{L}_{\text{NLL}}. \tag{6}$$

Algorithm 1 CoPO

Require: Preference dataset $\mathcal{D} = \{(x, y^+, y^-)\}$, reference model π_{ref}, policy π_θ

for each training step **do**
- Sample $(x, y^+, y^-) \sim \mathcal{D}$
- Sample anchor response $y^a \sim \pi_\theta(\cdot|x)$
- Compute implicit rewards of y^a, y^+, y^-: $R(x, y^a), R(x, y^+), R(x, y^-)$
- Compute normalized DPO Loss: $\mathcal{L}_{DPO} = -\log \sigma\left(\beta(R(x, y^+) - R(x, y^-))\right)$
- Compute reward trajectory alignment loss: $\mathcal{L}_{RTA} = \mathbb{E}_{t\sim T} - \log \frac{\exp(s\left(r_t^a, \text{sg}[r_t^+]\right)/\tau)}{\sum_{r'_t \in \{r_t^+, r_t^-\}} \exp(s(r_t^a, \text{sg}[r'_t])/\tau)}$
- Compute quality-aware weight w_i for the current preference sample
- Compute preferred-response regularization: $\mathcal{L}_{NLL} = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log \pi_\theta(y_t^+ | x, y_{<t})$
- Optimize: $\mathcal{L}_{CoPO} = \mathcal{L}_{DPO} + w_{RTA}\mathcal{L}_{RTA} + w_{NLL}\mathcal{L}_{NLL}$
- Update policy parameters by gradient descent: $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{CoPO}$

end for

return Optimized policy π_θ

3.5 Theoretical Analysis

Unlike standard contrastive learning that jointly updates positive and negative representations, CoPO performs asymmetric trajectory optimization. Reward trajectory alignment adopts stop-gradient on preference trajectories. Since stop-gradient removes optimization paths through preferred and rejected responses, i.e., $\frac{\partial L_{RTA}}{\partial r^+} = 0 \; and \; \frac{\partial L_{RTA}}{\partial r^-} = 0$, the optimization reduces to

$$\nabla_\theta \mathcal{L}_{RTA} = \frac{\partial \mathcal{L}_{RTA}}{\partial r^a} \frac{\partial r^a}{\partial \theta}. \tag{7}$$

Therefore, reward trajectory alignment updates only policy-generated anchor trajectories while keeping preference trajectories fixed. This asymmetric optimization differs fundamentally from conventional contrastive learning and can be interpreted as a projection process in implicit reward space.

Specifically, let $\mathcal{M}^+$ denote the manifold induced by preferred reward trajectories. Minimizing Eq.(4) approximately performs:

$$r^a \leftarrow \Pi_{\mathcal{M}^+}(r^a),$$

while maintaining separation from rejected trajectories. Unlike DPO, which enlarges reward margins only over observed responses, CoPO propagates preference supervision to policy-generated trajectories. Consequently, CoPO transforms preference optimization from $\max R(y^+) - R(y^-)$ into $\min D(r^a, \mathcal{M}^+)$, where generated behaviors are iteratively corrected toward preference-consistent reward regions. As training proceeds, newly generated anchors become increasingly aligned with preferred reward structures, expanding the coverage of preference-consistent trajectories.

4 Experiments

4.1 Experimental Setups

Evaluation Benchmarks. We evaluate model performance on seven benchmarks covering reasoning and knowledge understanding: MMLU (Hendrycks et al., 2021), which evaluates broad multitask knowledge and reasoning across diverse academic domains; ARC-Challenge (ARC-C) and ARC-Easy (ARC-E) (Clark et al., 2018), which measure scientific question answering ability at different difficulty levels; CommonsenseQA (Talmor et al., 2019), which tests commonsense reasoning; HellaSwag (Zellers et al., 2019), which evaluates grounded commonsense inference and sentence completion; TruthfulQA (Lin et al., 2022), which measures the ability to generate truthful responses and avoid common misconceptions; and Winograd Schema Challenge (WS) (Levesque et al., 2012), which assesses pronoun resolution and commonsense reasoning.

All evaluations employ greedy decoding and zero-shot chain-of-thought prompting for consistency. Following standard evaluation protocols, we use standard accuracy (`acc`) for MMLU, CommonsenseQA, and WS, normalized accuracy (`acc_norm`) for ARC-C, ARC-E, and HellaSwag, and MC2 accuracy (`acc_mc2`) for TruthfulQA.

Comparison Baselines. We compare CoPO with several preference optimization methods. Supervised-finetuning tuning corresponding to the **Base** model in our results, serves as the lower-bound reference. **DPO** (Rafailov et al., 2023) uses the log-likelihood ratio between the current policy and a reference model. **CPO** (Xu et al., 2024a) jointly optimizes preference and supervised objectives within a unified objective function. **SimPO** (Meng et al., 2024) is a reference-free method and uses length-normalized average log-likelihood as the implicit reward. **TIDPO** (Yang et al., 2026) estimates token-level importance using gradient attribution and a Gaussian prior, together with a triplet loss for token-level preference supervision. **UniDPO** (Peng et al., 2026) uses dual weighting mechanisms based on expert score margins and focal loss to dynamically reweight preference pairs.

Models and Training Settings. We perform preference optimization on two LLM backbone models, i.e., Llama-3.1-8B-Instruct (AI@Meta, 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024). For efficient fine-tuning, we adopt parameter-efficient fine-tuning using LoRA (Hu et al., 2022). Specifically, we set the LoRA rank to $r = 64$, the scaling factor to $\alpha = 128$, and the dropout rate to 0.05. The target modules include `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The experiments employ UltraFeedback dataset (Cui et al., 2024), comprising 60,000 samples.

Following the instruction-tuning setup, we use the off-the-shelf instruction-tuned model as the initialization and regenerate the chosen and rejected response pairs with the base model. Specifically, for each prompt, we sample five responses using the base model with a sampling temperature of 0.8. We score them using a reward model (e.g., ArmoRM (Wang et al., 2024) and GPT-4o (Hurst et al., 2024)), and select the highest-scoring response as the chosen response y^+ and the lowest-scoring response as the rejected response y^-.

Implementations Hyperparameters. All experiments are conducted on 4 $\times$ NVIDIA A800-80GB GPUs. We train all models for one epoch with a per-device training batch size of 4 and gradient accumulation steps of 16. The maximum sequence length is set to 512, and the maximum prompt length is set to 256. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a cosine learning rate schedule and 10% warmup steps. The learning rate is set to 1.0×10^{-4}. We use bfloat16 training, gradient checkpointing, and ZeRO-2 implementation for memory-efficient training. For CoPO, we set the trajectory alignment coefficient to $\lambda_{\text{RTA}} = 0.01$

Method	MMLU	ARC-C	ARC-E	CommonQA	HellaSwag	TruthfulQA	WS	Avg.
backbone: Llama-3.1-8B-Instruct								
Base	68.43	55.29	79.92	76.41	79.53	54.58	73.88	69.72
DPO	**68.52**	56.57	80.05	76.49	80.27	55.80	74.43	70.31
CPO	67.86	56.23	79.92	**77.23**	79.20	59.69	74.66	70.68
SimPO	67.12	65.27	80.60	76.41	75.07	67.84	78.14	72.92
TIDPO	67.02	64.25	79.55	75.18	73.97	64.65	76.48	71.59
UniDPO	68.12	64.93	**85.52**	76.90	83.95	68.20	78.69	75.19
CoPO (Ours)	67.49	**67.24**	**85.52**	76.74	**85.55**	**71.51**	**80.35**	**76.34**
backbone: Qwen-2.5-7B-Instruct								
Base	71.68	55.12	81.36	82.80	80.55	64.62	71.11	72.46
DPO	**71.73**	58.02	81.86	81.74	81.54	65.37	70.96	73.03
CPO	71.25	51.37	76.98	82.80	79.05	57.71	73.48	70.38
SimPO	71.55	61.60	66.58	82.80	67.97	64.99	67.48	69.00
TIDPO	71.64	57.34	79.97	81.08	81.03	67.76	70.32	72.73
UniDPO	71.30	65.53	84.09	81.98	**83.81**	70.00	68.67	75.05
CoPO (Ours)	71.26	**66.30**	**84.64**	**83.46**	82.47	**70.49**	**71.35**	**75.71**

Table 1: Average scores of each fine-tuning method with different LLM backbones.

Method	MMLU	ARC-C	ARC-E	CommonQA	HellaSwag	TruthfulQA	WS	Avg.
backbone: Llama-3.1-8B-Instruct								
Base	**68.43**	55.29	79.92	76.41	79.53	54.58	73.88	69.72
CoPO	67.49	67.24	85.52	**76.74**	**85.55**	**71.51**	**80.35**	**76.34**
w/o $\mathcal{L}_{\text{RTA}}$	68.20	65.27	83.75	76.66	84.61	68.38	78.93	75.11
w/o $\mathcal{L}_{\text{NLL}}$	67.51	**68.43**	**86.53**	76.49	85.04	68.05	79.01	75.87
w/o $\mathcal{L}_{\text{RTA}}$ & $\mathcal{L}_{\text{NLL}}$	67.50	66.04	83.46	76.25	83.71	67.01	78.37	74.62
backbone: Qwen-2.5-7B-Instruct								
Base	71.68	55.12	81.36	82.80	80.55	64.62	71.11	72.46
CoPO	71.26	**66.30**	**84.64**	**83.46**	82.47	**70.49**	71.35	**75.71**
w/o $\mathcal{L}_{\text{RTA}}$	**71.72**	60.84	82.11	82.31	**83.84**	67.64	**72.22**	74.38
w/o $\mathcal{L}_{\text{NLL}}$	71.40	63.48	70.08	83.13	72.10	65.68	66.85	70.39
w/o $\mathcal{L}_{\text{RTA}}$ & $\mathcal{L}_{\text{NLL}}$	70.96	63.40	69.49	83.29	71.22	64.13	68.11	70.09

Table 2: Ablation results of our method with different LLM backbones.

and the contrastive temperature to $\tau = 0.2$. We use the sample-anchor strategy to construct policy anchors. Anchor responses are sampled with temperature 0.7 and top-$p = 0.9$, and the maximum number of newly generated anchor tokens is set to $K = 8$. For the DPO-style objective, we set $\beta = [2, 0.1]$. The coefficient of the NLL calibration term is set to $\lambda_{\text{NLL}} = 0.01$.

4.2 Main Results

Table 1 presents the main results of CoPO and representative preference optimization baselines across reasoning and knowledge benchmarks. Overall, CoPO achieves better performance among all compared methods, demonstrating the effectiveness of our method. Moreover, CoPO shows consistent improvements on different LLM backbones, indicating that our method's robustness to different backbones. Specifically, CoPO improves over DPO by **+6.6**% and **+3.3**% in terms of average performance under Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, respectively.

4.3 Ablation Study

To understand the contribution of each component in CoPO, we conduct ablation experiments by removing the key objective in $\mathcal{L}_{\text{CoPO}}$ on two LLM backbones. The results are reported in Table 2. We compare the full CoPO objective with several variants by removing the reward trajectory alignment loss $\mathcal{L}_{\text{RTA}}$, the negative log-likelihood calibration term $\mathcal{L}_{\text{NLL}}$, or both of them. First, removing $\mathcal{L}_{\text{RTA}}$ weakens the overall effectiveness of CoPO. This shows that the reward trajectory alignment term is the key component to help the model better capture fine-grained reward evolution during generation. Second, removing $\mathcal{L}_{\text{NLL}}$ leads to inferior results, showing its necessity. $\mathcal{L}_{\text{NLL}}$ mainly serves as an auxiliary calibration term to stabilize trajectory alignment. When $\mathcal{L}_{\text{RTA}}$ and $\mathcal{L}_{\text{NLL}}$ are removed, the model performs worse than full CoPO, confirming that the two components are complementary.

4.4 Analysis of Anchor Response Length

We further evaluate the effect of the anchor response length K. The anchor length controls how

Method	MMLU	ARC-C	ARC-E	CommonQA	HellaSwag	TruthfulQA	WS	Avg.
Base	**68.43**	55.29	79.92	76.41	79.53	54.58	73.88	69.72
CoPO								
$K = 0$	68.07	66.30	84.43	75.27	84.59	66.05	78.37	74.73
$K = 4$	68.42	66.81	84.76	76.00	84.67	66.94	78.22	75.12
$K = 8$	67.49	**67.24**	**85.52**	76.74	**85.55**	**71.51**	**80.35**	**76.34**
$K = 16$	67.60	66.55	85.06	**76.82**	85.18	68.26	77.82	75.33

Table 3: Results (%) against lengths of anchor responses. We use Qwen-2.5-7B-Instruct as the backbone model.

many generated tokens are used to construct the policy anchor trajectory. We vary K in $\{0, 4, 8, 16\}$, where $K = 0$ means that we do not generate an anchor response and instead use the output distribution at the last prompt token as the anchor representation.

The results against different K on seven benchmarks are shown in Table 3. Compared with $K = 0$, using generated anchor responses $K > 0$ consistently improves most benchmarks. This shows that relying only on the prompt-end output provides limited behavioral information, while generated anchor tokens can better reflect the reward evolution of the current policy during autoregressive decoding. When K is small such as $K = 4$, the method only captures a short partial trajectory. Although it already improves over $K = 0$ on several benchmarks, the limited generation context is still insufficient to fully characterize the model's response behavior. Increasing K provides richer trajectory information and leads to the best overall result when $K = 8$. Further increasing K to 16 does not bring additional gains. This suggests that overly long anchors may introduce unnecessary complexity and accumulate noise from later-stage generation, where tokens can become redundant or less preference-relevant.

4.5 Generalization Evaluation on Multi-Backbone Preference Data

To mitigate the distribution shift models and the preference optimization process, we evaluate the generalization ability of our method on preference dataset generated using multiple LLM backbones rather than the target model itself. Given the prompts in UltraFeedback, we use preference responses generated with 17 LLMs. The results of our method and three baselines trained on the multi-backbone preference data on four benchmarks are shown in Figure 3. CoPO maintains strong performance and consistently outperforms baselines. This indicates that CoPO does not simply overfit to the response distribution of a particular generator. Instead, its effectiveness comes from using the current policy's anchor trajectory to connect external preference supervision with the model's own generation behavior. The reward trajectory alignment provides a robust mechanism for transferring preferences across different response distributions, demonstrating the generalization ability of CoPO.

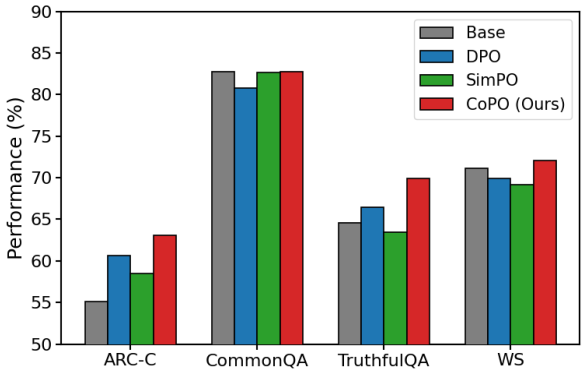


Figure 3: Results (%) of our method and baseline methods trained on preference data generated multiple LLM backbones. We use Llama-3.1-8B-Instruct as the backbone.

5 Related Work

Aligning large language models (LLMs) with human preferences has become a standard paradigm in post-training. Existing alignment methods can generally be divided into *explicit reward optimization* and *preference optimization*.

Explicit Reward Optimization Explicit reward optimization methods, such as reinforcement learning from human feedback (RLHF), learn reward functions and optimize policies through reinforcement learning. Representative approaches including PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and related variants (Zheng et al., 2025; Gao et al., 2025; Xu and Ding, 2026; Qi et al., 2026) achieve strong alignment performance but often require additional reward modeling and iterative policy updates.

Preference Optimization Preference optimization methods replace explicit reward estimation with direct policy optimization over preference pairs. Direct Preference Optimization (DPO) (Rafailov et al., 2023) reformulates preference optimization as implicit reward maximization and derives a closed-form objective without reinforcement learning. Subsequent methods extend this formulation from different perspectives. SimPO (Meng et al., 2024) removes the dependency on reference models through normalized likelihood optimization. Some methods exploit richer supervision structures (Xu et al., 2024b,a). IPO (Azar et al., 2024) revisits preference learning from an implicit reward perspective, while KTO (Ethayarajh et al., 2024) extends optimization beyond pairwise preference supervision. R-DPO (Park et al., 2024) introduces regularization to mitigate undesired length bias, and Uni-DPO (Peng et al., 2026) improves preference utilization through adaptive quality-aware weighting.

Finegrained Preference Optimization Recent works improve finer-grained preference supervision beyond sequence-level objectives. Token-aware preference optimization methods (Zeng et al., 2024; Zhu et al., 2025; Liu et al., 2025) extend preference learning to local token rewards and improve preference estimation at intermediate decoding stages. Trajectory-based methods (Yang et al., 2026) further incorporate generated responses into optimization and introduce additional structural constraints over generated outputs.

However, existing methods typically optimize absolute reward distances or complete response objectives. In contrast, CoPO preserves token-level reward trajectories and performs relative trajectory alignment through contrastive optimization over implicit rewards. Rather than enforcing explicit reward distance constraints, CoPO aligns sampled reward trajectories toward preferred behaviors through normalized preference signals.

6 Conclusion

We identify a fundamental limitation of existing preference optimization methods: optimization mismatch between preference objectives and autoregressive generation, which may progressively lead to preference collapse. To address this issue, we propose Contrastive Preference Optimization (CoPO), which aligns preference supervision with generation behavior through reward trajectory alignment over policy-generated anchors. By contrastively pulling generated trajectories toward preferred responses and separating them from rejected ones, CoPO encourages broader preference-consistent generation. Experiments across seven benchmarks and multiple LLM backbones demonstrate consistent improvements and strong generalization ability.

Limitations

Our work has several limitations. First, experiments are mainly conducted on public HuggingFace benchmarks and offline preference datasets, which may not fully reflect real-world deployment scenarios. Second, CoPO is only evaluated in the offline preference optimization setting. Although anchor responses partially reduce the mismatch between training and inference behaviors, we do not study online preference collection or iterative alignment. Third, due to computational constraints, experiments are limited to moderate-scale open-source LLMs. Whether CoPO scales consistently to larger frontier models remains for future investigation.

References

AI@Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Zhaohan Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2024. Implicit preference optimization. *arXiv preprint arXiv:2405.14734*.

Peter Clark and 1 others. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9722–9744.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, pages 12634–12651. PMLR.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. 2025. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*.

Dan Hendrycks and 1 others. 2021. Measuring massive multitask language understanding. *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. *KR*, 2012(13th):3.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics*, pages 3214–3252.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, and 1 others. 2025. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. In *International Conference on Learning Representations*, volume 2025, pages 51339–51368.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 124198–124235.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017.

S. Peng, W. Wang, Z. Tian, S. Yang, X. Wu, H. Xu, C. Zhang, T. Isobe, B. Hu, and M. Zhang. 2026. Uni-dpo: A unified paradigm for dynamic preference optimization of llms. In *The Fourteenth International Conference on Learning Representations*.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. 2026. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55204–55224.

J. Xu, A. Lee, S. Sukhbaatar, and J. Weston. 2024b. Swepo: Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Zhongwen Xu and Zihan Ding. 2026. Single-stream policy optimization. In *The Fourteenth International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. 2026. Token-importance guided direct preference optimization. In *The Fourteenth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *Proceedings of the*

41st International Conference on Machine Learning, pages 58348–58365.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025. Tgdpo: Harnessing token-level reward guidance for enhancing direct preference optimization. In *International Conference on Machine Learning*, pages 79727–79745. PMLR.