Mitigating Mismatch within Reference-based Preference Optimization

Suqin Yuan, The University of Sydney
Xingrui Yu, A*STAR CFAR
Jiyang Zheng, The University of Sydney
Lei Feng, Southeast University
Dadong Wang, CSIRO, Data61
Ivor Tsang, CFAR, A*STAR
Tongliang Liu, The University of Sydney

The One-Line Change for DPO Loss Function: Replace Δθ – Δref with Δθ –max(0, Δref)

ICLR

DPO optimizes the **relative margin** $\Delta_\theta - \Delta_{ref}$ between the policy and a reference $\Delta_\theta - \Delta_{ref}$ to keep updates within a stable "trusted region".

- Trainable policy margin: $\Delta_\theta = \log \pi_\theta(y^+|x) - \log \pi_\theta(y^-|x)$
- Frozen reference margin: $\Delta_{ref} = \log \pi_{ref}(y^+|x) - \log \pi_{ref}(y^-|x)$

Premature Satisfaction — occurs on data pairs where the reference model prefers the rejected response y^-. **Issue**: The gradient decays as soon as the policy simply outperforms the reference on this pair, even if the policy is still wrong $\Delta_\theta < 0$.

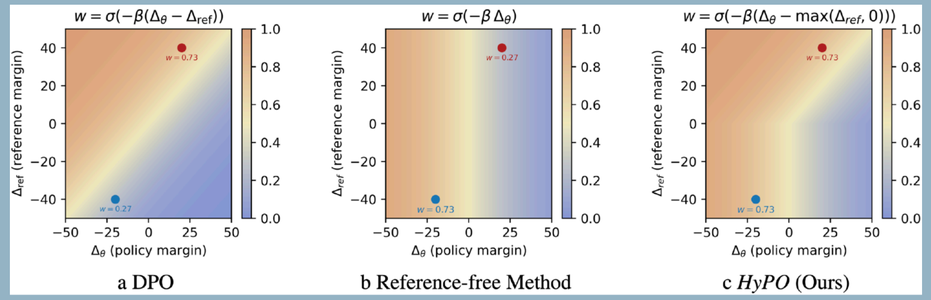
Reference-free Objectives, directly optimize the policy margin Δ_θ

Without the reference policy Δ_{ref} "trusted region", for pairs y^+ and y^- from different distributions, it potentially over-optimization and loses basic linguistic ability.

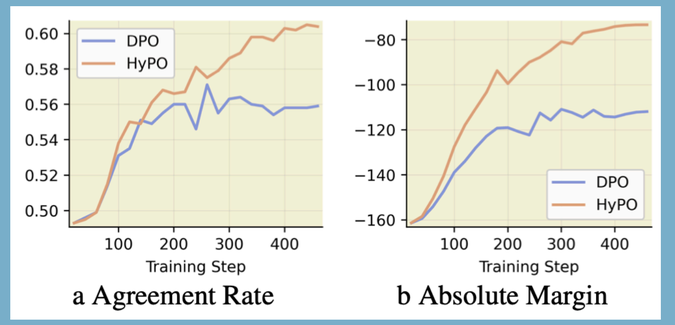
Our Solution: **Hybrid-DPO (HyPO)**, conditionally apply the reference:

If $\Delta_{\text{ref}} \geq 0$: Behaves exactly like standard DPO to preserve the "trusted region" stability.

If $\Delta_{\text{ref}} < 0$: Ignores the misleading reference, use absolute margin Δ_θ to optimize hardest examples.



a DPO b Reference-free Method c *HyPO* (Ours)



a Agreement Rate b Absolute Margin

- HyPO bridges the training-inference gap, achieving a faster and higher increase in both Agreement Rate and Absolute Margin.

Table 1: Main results reported from AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard-v0.1 (Li et al., 2024a;b). LC and WR denote *length-controlled win rate* and *raw win rate*, respectively. The best results are highlighted in bold, and the second-best are underlined.

Method	Mistral-Base (7B)			Mistral-Instruct (7B)		
	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard
	LC (%)	WR (%)	WR (%)	LC (%)	WR (%)	WR (%)
SLiC-HF (Zhao et al., 2023)	11.6	9.1	5.4	32.4	31.2	16.9
DPO (Rafailov et al., 2023)	22.6	18.5	7.9	35.1	31.4	15.4
CPO (Xu et al., 2024)	13.1	11.6	6.4	34.9	39.9	<u>21.0</u>
KTO (Ethayarajh et al., 2024)	12.9	9.3	6.6	35.0	31.3	17.5
SimPO (Meng et al., 2024)	27.3	25.4	<u>11.2</u>	<u>38.4</u>	<u>40.0</u>	20.5
FocalPO (Liu et al., 2025)	25.8	19.7	8.2	35.9	35.0	18.7
TR-DPO (Gorbatovski et al., 2025)	24.9	21.4	9.5	36.5	33.7	18.2
RainbowPO (Zhao et al., 2025)	<u>28.4</u>	<u>26.7</u>	9.2	35.7	33.9	18.2
HyPO (Ours)	**32.8**	**29.6**	**13.9**	**38.9**	**47.9**	**25.2**
Method	**Llama-3-Base (8B)**			**Llama-3-Instruct (8B)**		
	AlpacaEval		Arena-Hard	AlpacaEval		Arena-Hard
	LC (%)	WR (%)	WR (%)	LC (%)	WR (%)	WR (%)
SLiC-HF (Zhao et al., 2023)	19.8	15.9	14.3	36.7	36.8	25.1
DPO (Rafailov et al., 2023)	24.3	21.9	23.0	40.9	41.3	31.5
CPO (Xu et al., 2024)	22.3	24.6	12.2	38.1	40.4	30.0
KTO (Ethayarajh et al., 2024)	23.6	20.3	18.4	40.5	39.0	30.5
SimPO (Meng et al., 2024)	30.7	26.2	30.1	46.0	43.1	32.1
FocalPO (Liu et al., 2025)	27.2	25.4	27.9	45.1	<u>43.6</u>	30.2
TR-DPO (Gorbatovski et al., 2025)	<u>31.8</u>	<u>30.2</u>	<u>31.0</u>	<u>46.7</u>	42.7	<u>32.5</u>
RainbowPO (Zhao et al., 2025)	30.3	27.1	28.6	<u>46.7</u>	43.5	31.3
HyPO (Ours)	**34.7**	**33.6**	**33.5**	**49.5**	**46.2**	**35.2**

- Consistently dominates baselines, yielding a 41.2% average relative improvement over standard DPO.
- Introduces zero extra overhead and can be combined with other DPO improvements, such as average log-probs, home advantage margin or stronger reference model.

1.

2.