

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Author: Rafailov et al.

created by paper2slides

Uploaded: 2023-05-29

Executive Summary

- Language models (LMs) trained unsupervised lack precise control due to the diverse nature of the data.
- RLHF is complex and unstable; it requires fitting a reward model and fine-tuning LMs.
- The paper proposes *Direct Preference Optimization (DPO)*: a method to directly train LMs on human preferences.
- DPO leverages a mapping between reward functions and optimal policies, simplifying preference learning without RL.
- Experimental results show DPO outperforms or matches PPO-based RLHF on several tasks.

Introduction

- Large unsupervised LMs possess broad knowledge and reasoning skills, but task-specific control remains challenging.
- Current methods align LMs using human preferences through RLHF, which involves complex and unstable procedures.
- Reinforcement Learning (RL) methods typically require multiple stages:
 - Training a reward model reflecting human preferences.
 - Fine-tuning the LM using RL to maximize the reward.
- This paper introduces a more efficient alternative: Direct Preference Optimization (DPO).

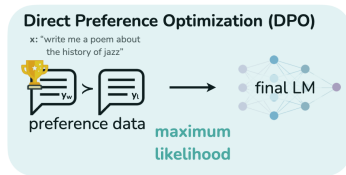
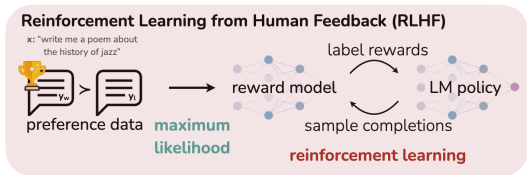


Figure: DPO optimizes for human preferences directly, avoiding RL by reparameterizing the reward function.

Direct Preference Optimization (DPO)

- Key insight: Use a theoretical mapping from reward functions to optimal policies.
- This allows transforming loss function over reward functions into a loss function over policies.
- Formulate objectives directly on policies using equations from the Bradley-Terry model.

Derivation of DPO Objective

$$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x) \exp \left(\frac{1}{\beta} r(x, y) \right)$$

- π_r : optimal policy, π_{ref} : reference policy, r : reward function, β : KL-divergence constraint.

$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

- Reformulate r in terms of π_r and π_{ref} , enabling direct policy optimization.

DPO Objective based on Bradley-Terry Model

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

- σ : logistic function, π_{θ} : parameterized policy.
- Leverages human preference data directly for optimization.

Theoretical Analysis of DPO

- DPO avoids explicit reward estimation and RL.
- Proves capability to represent all reward functions equivalent under Plackett-Luce model.
- Offers a reliable optimization procedure compared to actor-critic algorithms.
- Relevant Theorem:
 - All rewards consistent with the Bradley-Terry model can be expressed as $\beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}.$

Empirical Evaluation: Sentiment Generation

- Task: Generate positive sentiment movie reviews.
- Evaluation: Reward-KL frontier using a sentiment classifier as the ground truth.
- Comparison: PPO, PPO-GT, Preferred-FT, Unlikelihood.
- Results:
 - DPO achieves highest reward for given KL values.
 - DPO strictly outperforms PPO and algorithms that access ground truth rewards (PPO-GT).

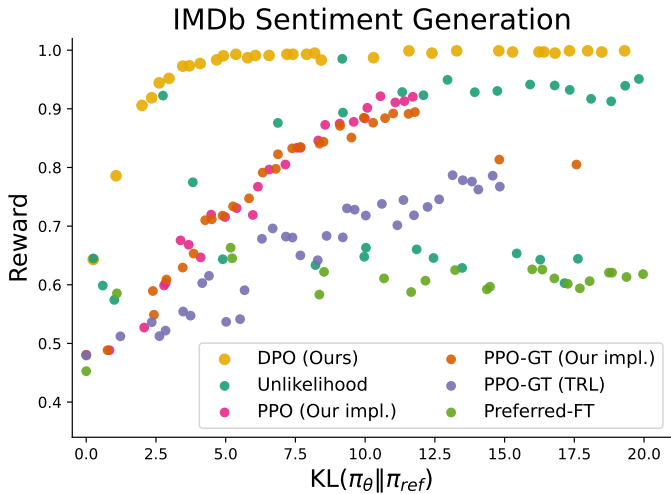


Figure: DPO provides the highest expected reward for all KL values, demonstrating efficient optimization.

Empirical Evaluation: Summarization

- Task: Summarize Reddit posts using TL;DR dataset.
- Evaluation: GPT-4 win rate against human-written summaries.
- Comparison: PPO, Preferred-FT, Best of N , Zero-shot GPT-J.
- Results:
 - DPO exceeds PPO's best-case performance.
 - DPO is more robust across different sampling temperatures.

TL;DR Summarization Win Rate vs Reference

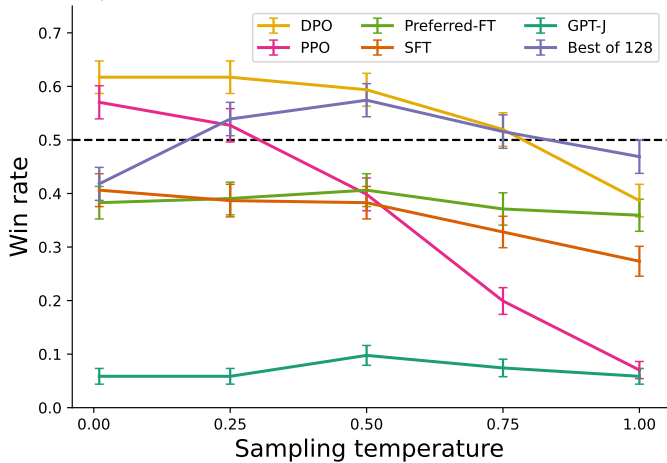


Figure: DPO exceeds PPO-based RLHF in summarization performance and robustness across temperatures.

Empirical Evaluation: Single-Turn Dialogue

- Task: Generate engaging and helpful responses to user queries using Anthropic HH dataset.
- Evaluation: GPT-4 win rate against preferred completions.
- Comparison: PPO, Preferred-FT, Best of N , Prompted Pythia-2.8B.
- Results:
 - DPO is the only method improving over chosen completions.
 - DPO provides similar or better performance than Best of 128.

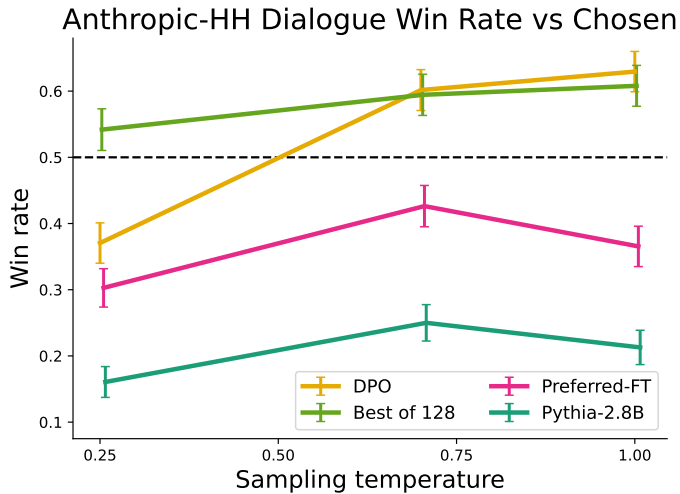


Figure: DPO is the only method that improves over chosen completions in the Anthropic HH dataset.

Generalization to New Input Distribution

- Task: Evaluate TL;DR models on CNN/DailyMail news articles.
- Results:
 - DPO continues to outperform PPO on out-of-distribution data.
- Indicates robust generalization capability of DPO policies.

Human Judgments Verification

- Human study to verify GPT-4 judgments.
- Metrics: Win rates and agreement rate with human evaluations.
- Results:
 - GPT-4's judgments correlate strongly with human judgments.
 - Human agreement with GPT-4 similar/higher than inter-human agreement.

Conclusion and Future Work

- DPO offers a simpler, stable alternative to RLHF by training LMs directly from human preferences.
- Achieves superior or competitive performance across multiple tasks.
- Future work: Explore further scaling to larger LMs, study OOD generalization, and identify methods to prevent reward over-optimization.