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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn

Stanford University

NeurIPS 2023 Outstanding Main Track Runner-Ups

Presenter: Hawon Jeong

https://arxiv.org/pdf/2305.18290.pdf

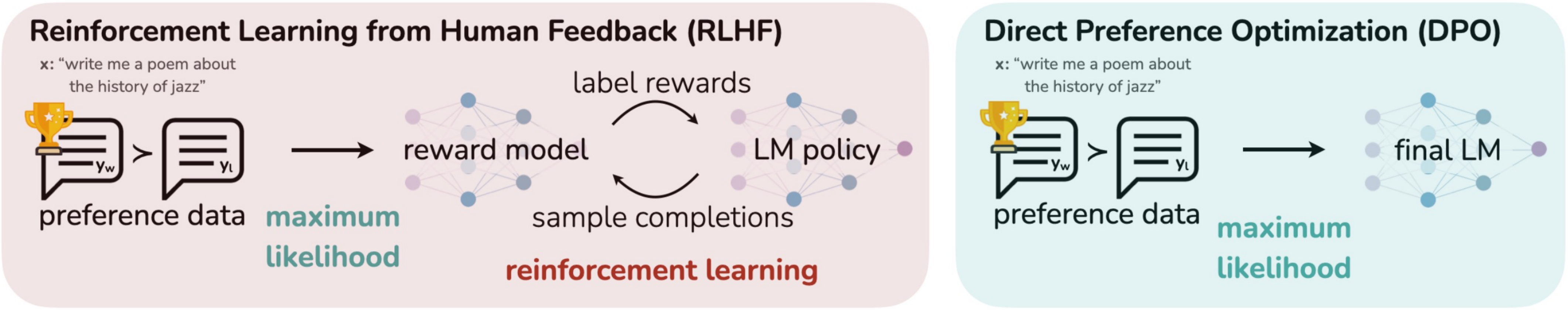
Introduction — Backgrounds

- LLMs are trained on data generated by humans with a wide variety of goals, priorities, and skillsets. Some of these goals and skillsets may not be desirable to imitate.
  - E.g., understanding common programming mistakes.

- Selecting the model's desired responses and behavior from its wide knowledge and abilities is crucial to building AI systems that are safe, performant, and controllable.

- The most straightforward approach to preference learning: Supervised Fine-Tuning
- The most successful class of methods: **RLHF**
  - Pros: Impressive conversational and coding abilities with **preference data**
  - Cons: Training multiple LMs, sampling from the LM policy in the loop of training …
  - **Complex** 😵😵😵

Introduction — DPO


Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
reward model
label rewards
sample completions
LM policy
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
final LM

- Goal: Directly optimizing a language model to adhere to human preference w/o explicit reward modeling or RL
- **Direct Preference Optimization (DPO)** implicitly optimizes the same objective as existing RLHF algorithms but **simple to implement and straightforward to train**
- Contribution
    - RL-free algorithm
    - Experiments show that DPO is as effective as existing methods, including PPO-based RLHF

Preliminaries — RLHF

SFT

- Fine-tuning a pre-trained LM with supervised learning on high-quality data for the downstream task(s) of interest

Reward Modeling Phase

- Human preference datasets: $\mathcal{D} = \left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$
- Human preference model(BT model): $p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$
- Estimation of reward model with maximum likelihood and NLL loss:

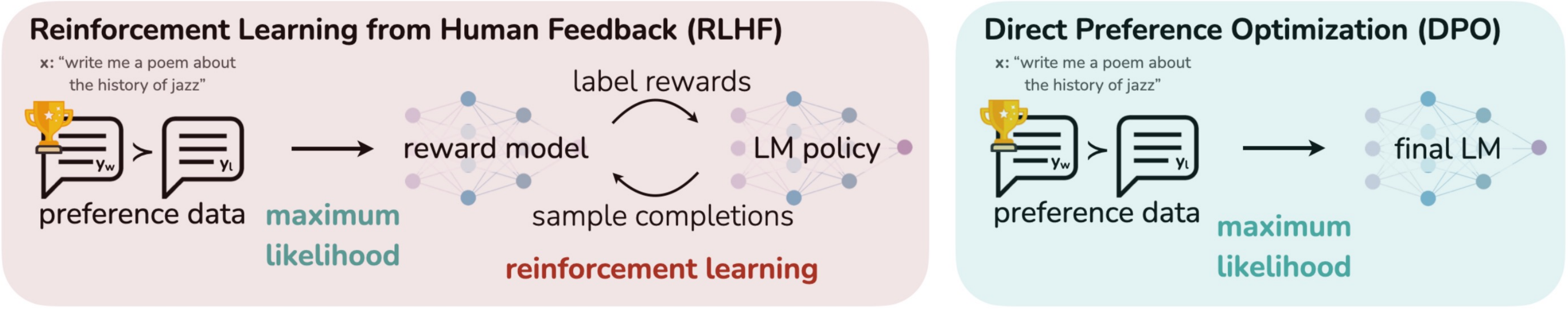
$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

RL Fine-Tuning Phase

- Optimization problem: $\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)}\left[r_\phi(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$
- Typically, maximize using PPO: $r(x, y) = r_\phi(x, y) - \beta(\log \pi_\theta(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x))$

https://arxiv.org/pdf/1909.08593.pdf

Direct Preference Optimization


Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
reward model
label rewards
sample completions
LM policy
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
final LM

- Our key insight is to leverage an analytical mapping from reward functions to optimal policies, which enables us to **transform a loss function over reward functions into a loss function over policies**.
- This change-of-variables approach avoids fitting an explicit, standalone reward model, while still optimizing under existing models of human preferences
- In essence, the policy network represents both the language model and the (implicit) reward.

Direct Preference Optimization — Deriving the DPO objective

- Change of variables

Target LM

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi}\big[r(x,y)\big] - \beta\mathbb{D}_{\text{KL}}\big[\pi(y|x) \mid\mid \pi_{\text{ref}}(y|x)\big]$$

Start with same RL objective as prior work

$$= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta}r(x,y)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right] \qquad (12)$$

where $\boxed{Z(x)} = \sum_{y}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right).$

Partition function which only depends on π_{ref} and x

Let $\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$ ← Probability distribution

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi^*(y|x)}\right] - \log Z(x)\right] = \qquad (13)$$

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{D}_{\text{KL}}(\pi(y|x) \mid\mid \pi^*(y|x)) - \log Z(x)\right] \qquad (14)$$

Optimal solution: $\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right) \qquad (15)$

Direct Preference Optimization — Deriving the DPO objective

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

Still expensive to estimate the partition function $Z(x)$

- Rearrange the equation to express the reward function in terms of π_r, π_{ref} and $Z(\cdot)$.

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

- The optimal RLHF policy π^* under BT model satisfies the preference model

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1+\exp\left(\beta\log\frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta\log\frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

BT model:

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x,y_1)\right)}{\exp\left(r^*(x,y_1)\right)+\exp\left(r^*(x,y_2)\right)}.$$

- Maximum likelihood objective for a parameterized policy π_θ

$$\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]. \qquad (7)$$

Reward model loss:

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

Direct Preference Optimization — What does DPO updates do?

- Gradient

$$\nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) =$$

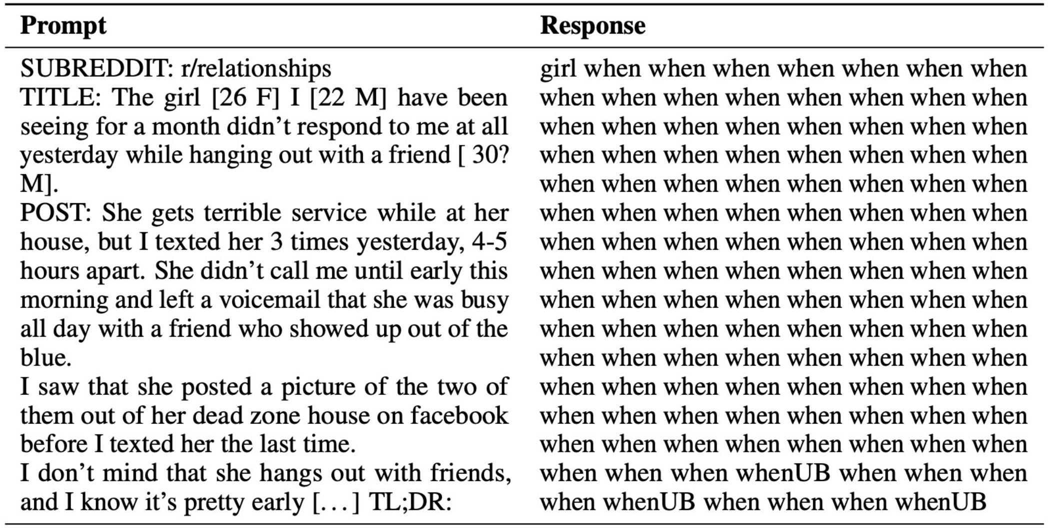
$$-\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\underbrace{\sigma(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w))}_{\text{higher weight when reward estimate is wrong}}\left[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l}\right]\right]$$

$\hat{r}_\theta(x,y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}$: implicit reward defined by π_θ and π_{ref}

- Intuitively, the gradient of the loss function $\mathcal{L}_{\mathrm{DPO}}$
  - increases the likelihood of the preferred completions y_w
  - decreases the likelihood of the dispreferred completions y_l

- Importantly, the examples are weighed by how much higher the implicit reward model $\hat{r}_\theta$ rates the dispreferred completions, scaled by β, i.e, **how incorrectly the implicit reward model orders the completions**, accounting for the strength of the KL constraint.

- Unlikelihood samples from TL;DR prompts →



Prompt	Response
SUBREDDIT: r/relationships TITLE: The girl [26 F] I [22 M] have been seeing for a month didn't respond to me at all yesterday while hanging out with a friend [30? M]. POST: She gets terrible service while at her house, but I texted her 3 times yesterday, 4-5 hours apart. She didn't call me until early this morning and left a voicemail that she was busy all day with a friend who showed up out of the blue. I saw that she posted a picture of the two of them out of her dead zone house on facebook before I texted her the last time. I don't mind that she hangs out with friends, and I know it's pretty early [. . .] TL;DR:	girl whenUB when when when when whenUB when when when whenUB

Direct Preference Optimization — DPO outline

- The general DPO pipeline is as follows:

  1) Sample completions $y_1, y_2 \sim \pi_{ref}(\cdot\,|x)$for every prompt x, label with human preferences to construct the offline dataset of preferences $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$

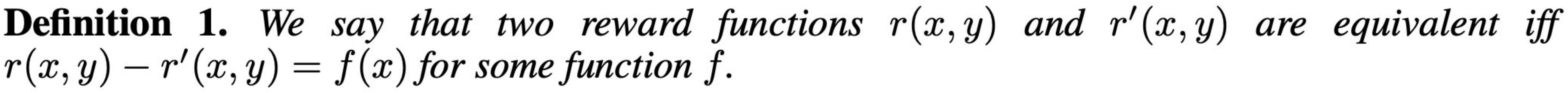
  2) optimize the language model π_θ to minimize $\mathcal{L}_{\text{DPO}}$ for the given π_{ref} and $\mathcal{D}$ and desired β

- Generally π_{ref} is initialize as π_{SFT}

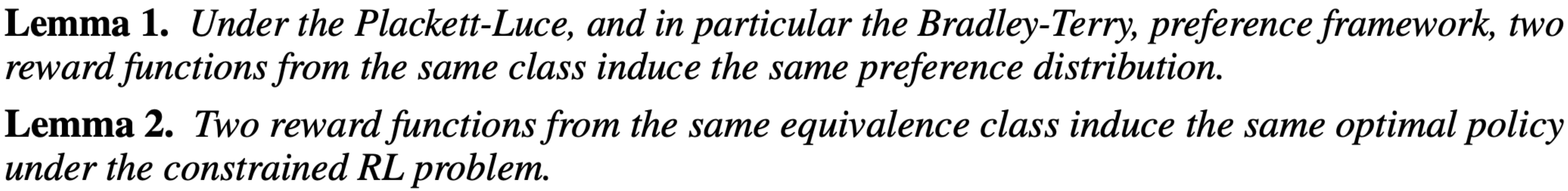
- However, when π_{SFT} is not available, we initialize π_{ref} by maximizing likelihood of preferred completions (x, y_w)

$$\pi_{\text{ref}} = \arg\max_\pi \mathbb{E}_{x,y_w\sim\mathcal{D}}\left[\log \pi(y_w \mid x)\right]$$

Theoretical Analysis — Your Language Model Is Secretly a Reward Model

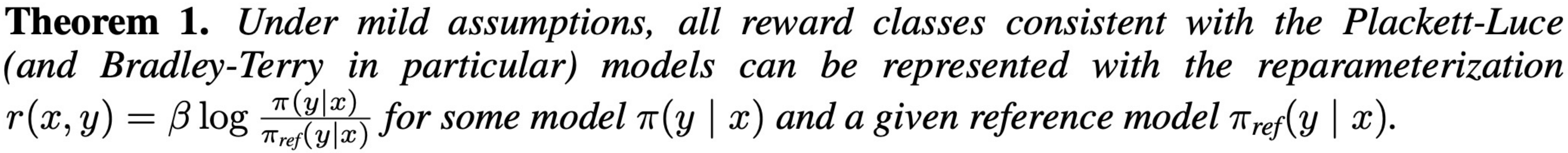


Definition 1. *We say that two reward functions* $r(x, y)$ *and* $r'(x, y)$ *are equivalent iff* $r(x, y) - r'(x, y) = f(x)$ *for some function* f.



Lemma 1. *Under the Plackett-Luce, and in particular the Bradley-Terry, preference framework, two reward functions from the same class induce the same preference distribution.*

Lemma 2. *Two reward functions from the same equivalence class induce the same optimal policy under the constrained RL problem.*



Theorem 1. *Under mild assumptions, all reward classes consistent with the Plackett-Luce (and Bradley-Terry in particular) models can be represented with the reparameterization* $r(x, y) = \beta \log \frac{\pi(y|x)}{\pi_{ref}(y|x)}$ *for some model* $\pi(y \mid x)$ *and a given reference model* $\pi_{ref}(y \mid x)$.

- We preserve the class of representable reward models, but explicitly make the optimal policy analytically tractable for all prompts x.

Theoretical Analysis — Instability of Actor-Critic Algorithms

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y\mid x) \mid\mid \pi_{\mathrm{ref}}(y\mid x)\right]$$

PPO objective

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta(y|x)}\Bigg[\underbrace{r_\phi(x,y) - \beta\log\sum_y \pi_{\mathrm{ref}}(y\mid x)\exp\left(\frac{1}{\beta}r_\phi(x,y)\right)}_{f(r_\phi,\pi_{\mathrm{ref}},\beta)} - \underbrace{\beta\log\frac{\pi_\theta(y\mid x)}{\pi_{\mathrm{ref}}(y\mid x)}}_{\mathrm{KL}}\Bigg]$$

PPO objective
+ DPO optimal policy π^*

- While the normalization term does not affect the optimal solution, without it, the policy gradient of the objective could have high variance, making learning unstable.

Source

Experiments — Experimental setup

- Controlled sentiment generation
  - x: prefix of a movie review from the IMDb dataset, y: positive sentiment
  - For SFT, we fine-tune GPT-2-large until convergence on reviews from the train split of the IMDB dataset

- Summarization
  - X: a forum post from Reddit(TL;DR), y: a summary of the main points in the post
  - We use an SFT model fine-tuned on human-written forum post summaries2 with the TRLX (https://huggingface.co/CarperAI/openai_summarize_tldr_sft)

- Single-turn dialogue
  - x: human query, y: engaging and helpful response y to a user's query
  - In this setting, no pre-trained SFT model is available; we therefore fine-tune an off-the-shelf language model on only the preferred completions to form the SFT model

Source

Experiments — How well can DPO optimize the RLHF objective?

- Sentiment Generation & Summarization

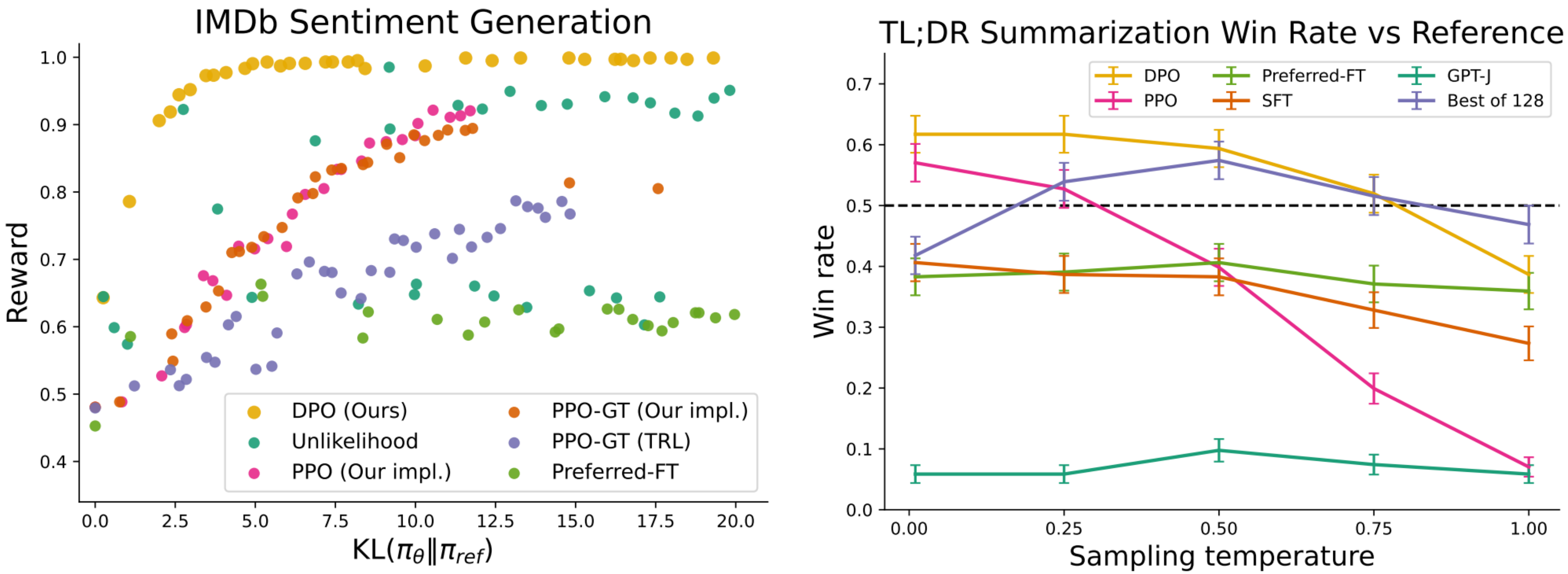


Figure 2: **Left.** The frontier of expected reward vs KL to the reference policy. DPO provides the highest expected reward for all KL values, demonstrating the quality of the optimization. **Right.** TL;DR summarization win rates vs. human-written summaries, using GPT-4 as evaluator. DPO exceeds PPO's best-case performance on summarization, while being more robust to changes in the sampling temperature.

Source

Experiments — Can DPO scale to real preference datasets?

- Dialogue

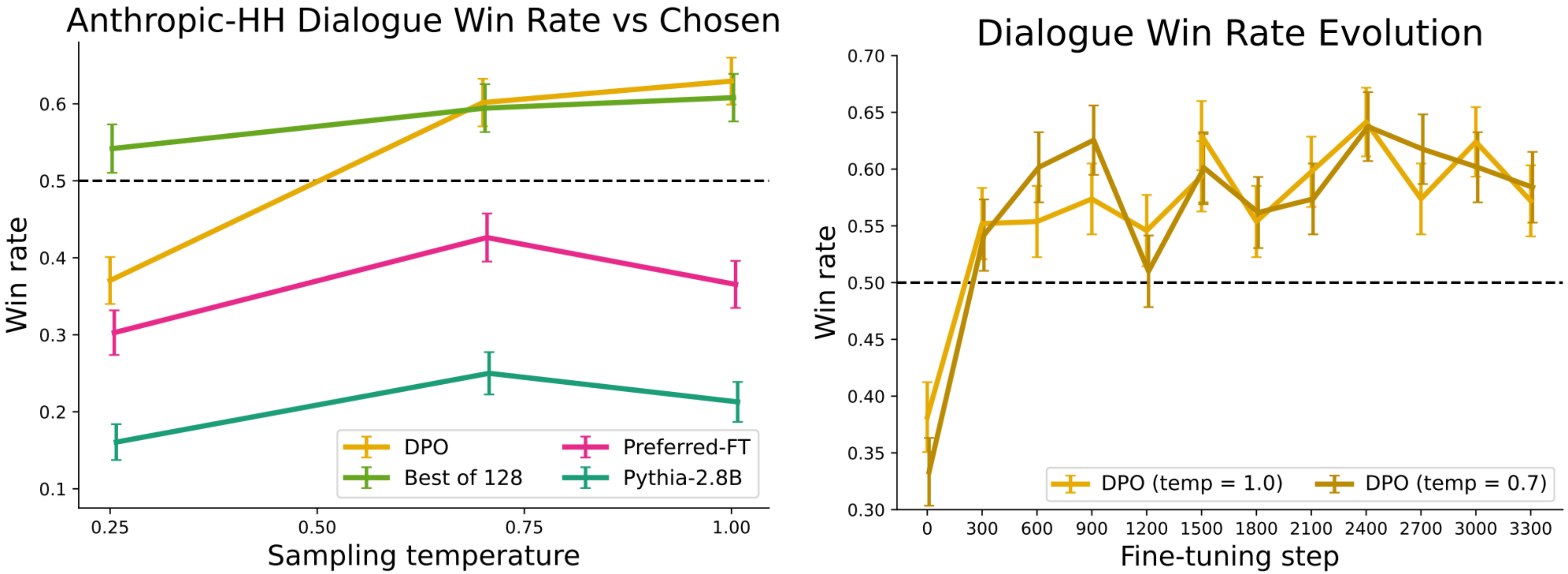
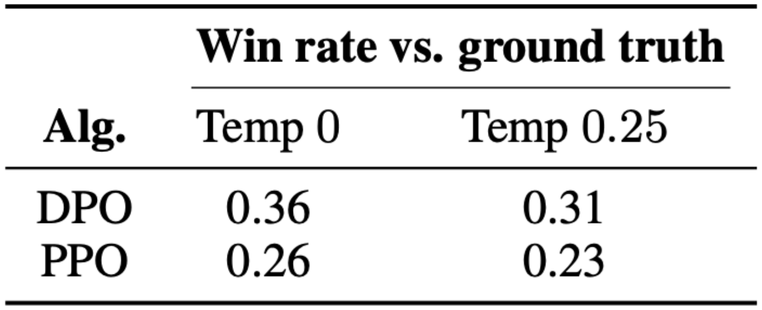


Figure 3: **Left.** Win rates computed by GPT-4 for Anthropic-HH one-step dialogue; DPO is the only method that improves over chosen summaries in the Anthropic-HH test set. **Right.** Win rates for different sampling temperatures over the course of training. DPO's improvement over the dataset labels is fairly stable over the course of training for different sampling temperatures.

Experiments — Generalization to a new input distribution

- Performance under distribution shift

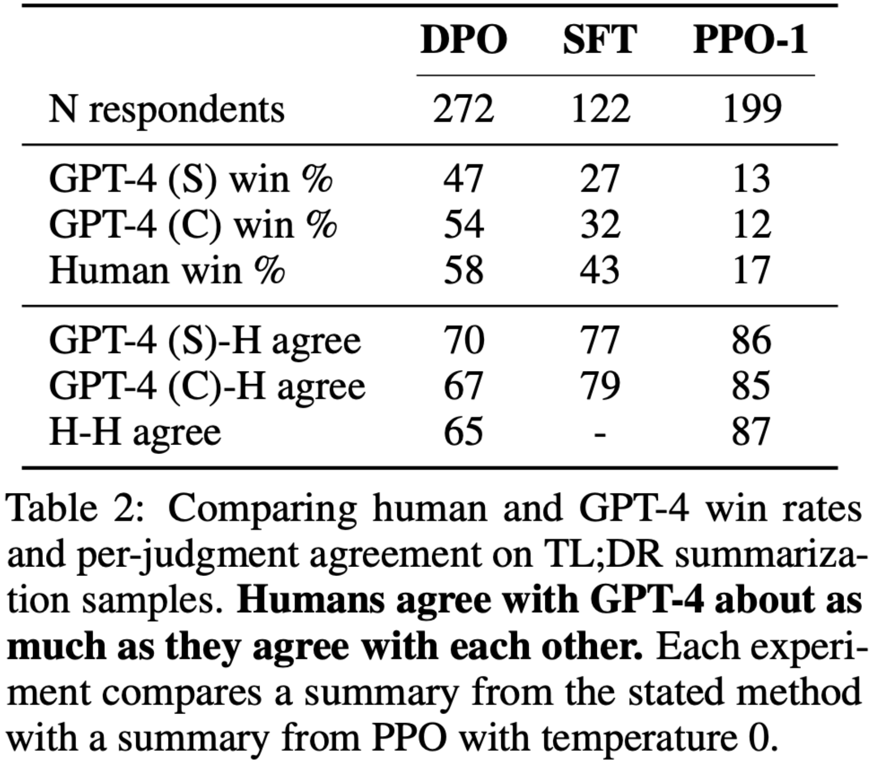


	Win rate vs. ground truth	
Alg.	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

Table 1: GPT-4 win rates vs. ground truth summaries for out-of-distribution CNN/DailyMail input articles.

Source

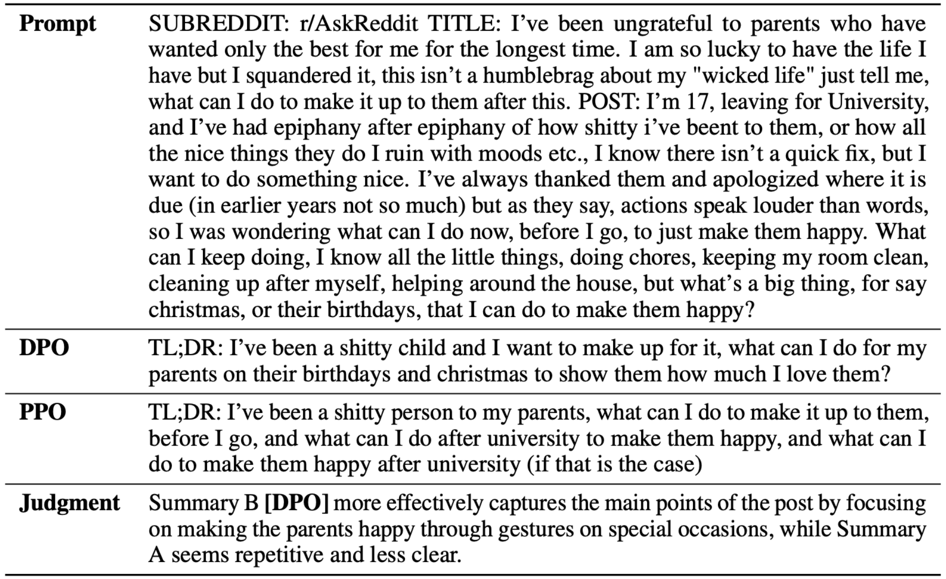
Experiments — Validating GPT-4 judgments with human judgments



	DPO	SFT	PPO-1
N respondents	272	122	199
GPT-4 (S) win %	47	27	13
GPT-4 (C) win %	54	32	12
Human win %	58	43	17
GPT-4 (S)-H agree	70	77	86
GPT-4 (C)-H agree	67	79	85
H-H agree	65	-	87

Table 2: Comparing human and GPT-4 win rates and per-judgment agreement on TL;DR summarization samples. **Humans agree with GPT-4 about as much as they agree with each other.** Each experiment compares a summary from the stated method with a summary from PPO with temperature 0.

- We find that GPT-4 prefers longer, more repetitive summaries than humans do with the GPT-4 (S) prompt
- Overall, GPT-4 is a reasonable proxy for human evaluations, providing win rates more representative of humans



Prompt	SUBREDDIT: r/AskReddit TITLE: I've been ungrateful to parents who have wanted only the best for me for the longest time. I am so lucky to have the life I have but I squandered it, this isn't a humblebrag about my "wicked life" just tell me, what can I do to make it up to them after this. POST: I'm 17, leaving for University, and I've had epiphany after epiphany of how shitty i've beent to them, or how all the nice things they do I ruin with moods etc., I know there isn't a quick fix, but I want to do something nice. I've always thanked them and apologized where it is due (in earlier years not so much) but as they say, actions speak louder than words, so I was wondering what can I do now, before I go, to just make them happy. What can I keep doing, I know all the little things, doing chores, keeping my room clean, cleaning up after myself, helping around the house, but what's a big thing, for say christmas, or their birthdays, that I can do to make them happy?
DPO	TL;DR: I've been a shitty child and I want to make up for it, what can I do for my parents on their birthdays and christmas to show them how much I love them?
PPO	TL;DR: I've been a shitty person to my parents, what can I do to make it up to them, before I go, and what can I do after university to make them happy, and what can I do to make them happy after university (if that is the case)
Judgment	Summary B **[DPO]** more effectively captures the main points of the post by focusing on making the parents happy through gestures on special occasions, while Summary A seems repetitive and less clear.

Discussion

- DPO identifies a mapping between language model policies and reward functions that enables training a language model to satisfy human preferences directly, with a simple cross-entropy loss, without reinforcement learning or loss of generality
- With virtually no tuning of hyperparameters, DPO performs similarly or better than existing RLHF algorithms, including those based on PPO

- Limitations & Future work
  - How does the DPO policy generalize out of distribution, compared with learning from an explicit reward function?
  - How does reward over-optimization manifest in the direct preference optimization setting?
    - Is the slight decrease in performance in Figure 3-right an instance of it?
  - Exploration of scaling DPO on larger model
  - Elicit high-quality judgments from automated systems
  - Applications of DPO beyond training language models from human preferences, including training generative models in other modalities