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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn
NeurIPS 2023

2025. 11. 13
Learning Agents 강화학습 논문 리뷰 스터디
Minkyoung Kim

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 [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Agenda

- Introduction
- Method
- Experiments
- Conclusion

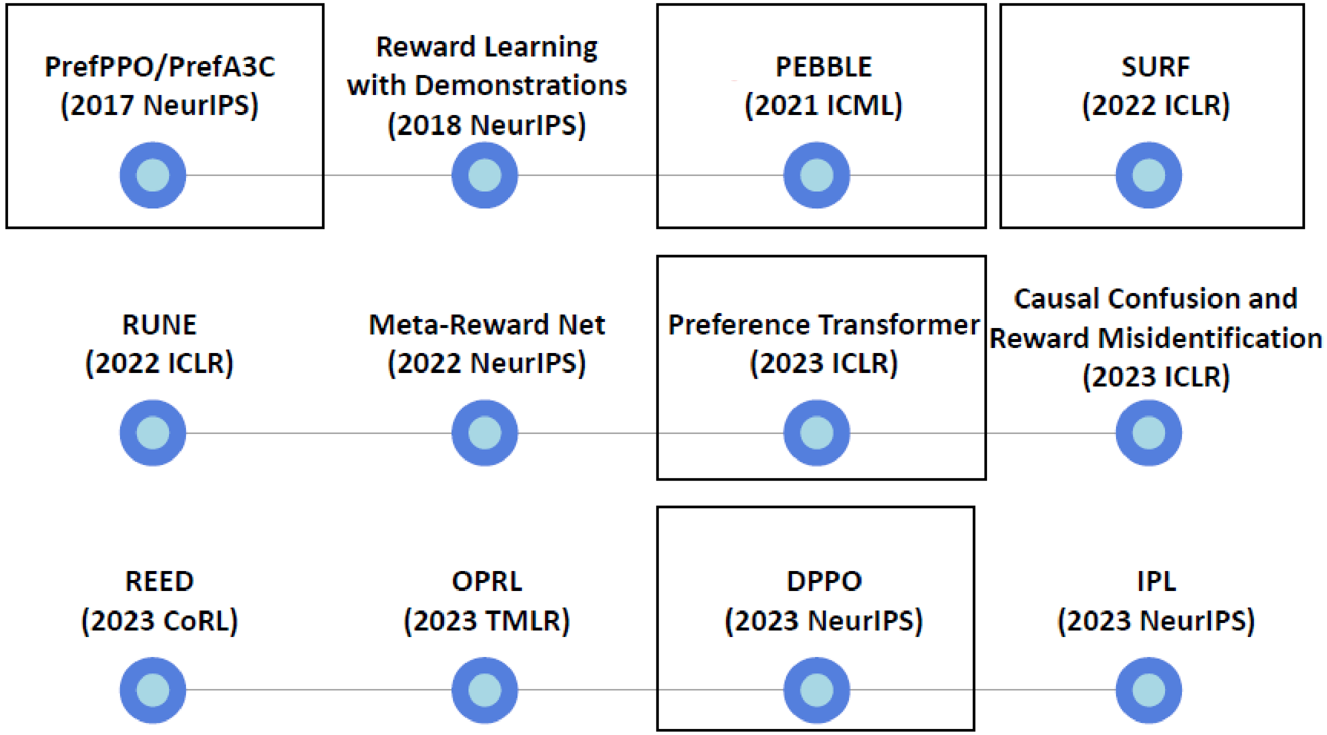
PrefPPO
- introduction of PbRL
- Reward Ensemble and Sampling
- on-policy Algorithm (PPO)

PEBBLE
- unsupervised Pre-training for Exploration
- off-policy Algorithm (SAC)
- Relabeling Replay Buffer for Stable Learning

SURF
- semi-supervised learning
- proposed data augmentation

Preference Transformer
- offline RL
- weighted sum of non-Markovian rewards

DPPO
- reward model-free, offline optimization

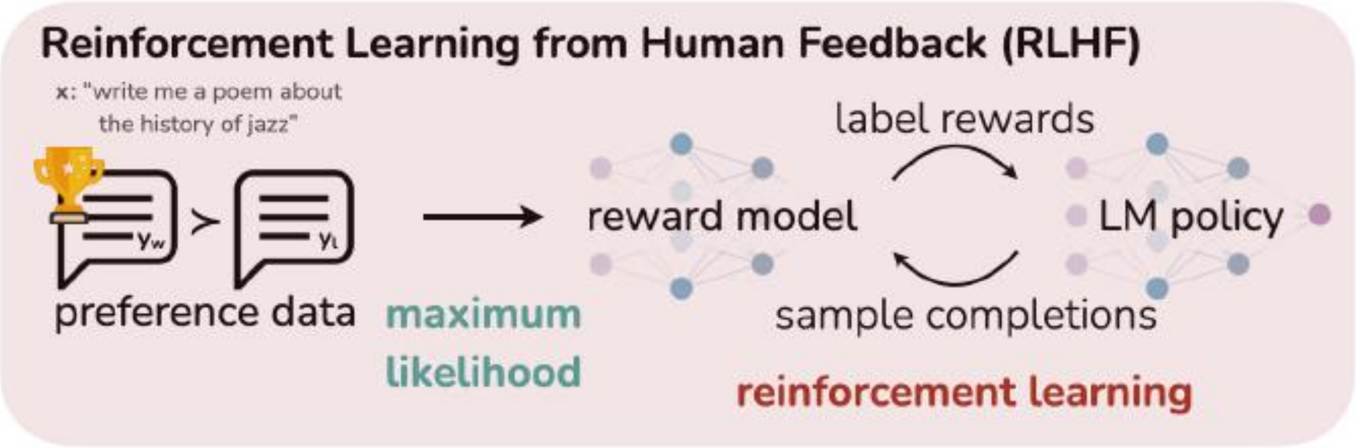


Introduction

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

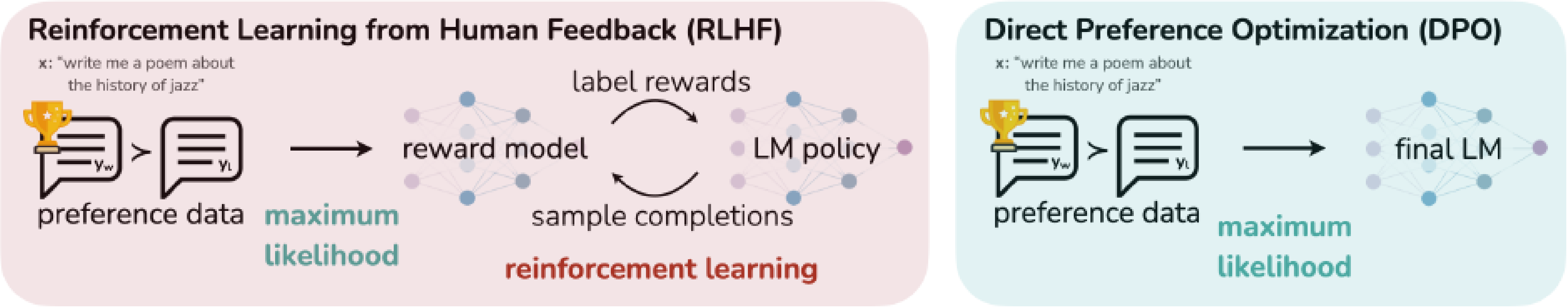
→ Steer LMs to match human preferences using reinforcement learning

- **RLHF** is a complex and often unstable procedure.
  - 1) fitting a reward model that reflects the human preferences
  - 2) fine-tuning the large unsupervised LM
    - to maximize this estimated reward without drifting too far from the original model



Introduction

- **Direct Preference Optimization(DPO)**
  - Optimize a LM to adhere to human preferences, **without explicit reward modeling or RL**
  - Uses a **change of variables** to define the preference loss as **function of policy** directly
  - stable, performant, and computationally lightweight,
    eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning



RLHF(Reinforcement Learning from Human Feedback)

1) Supervised fine-tuning (SFT)

2) preference sampling and reward learning

3) RL optimization

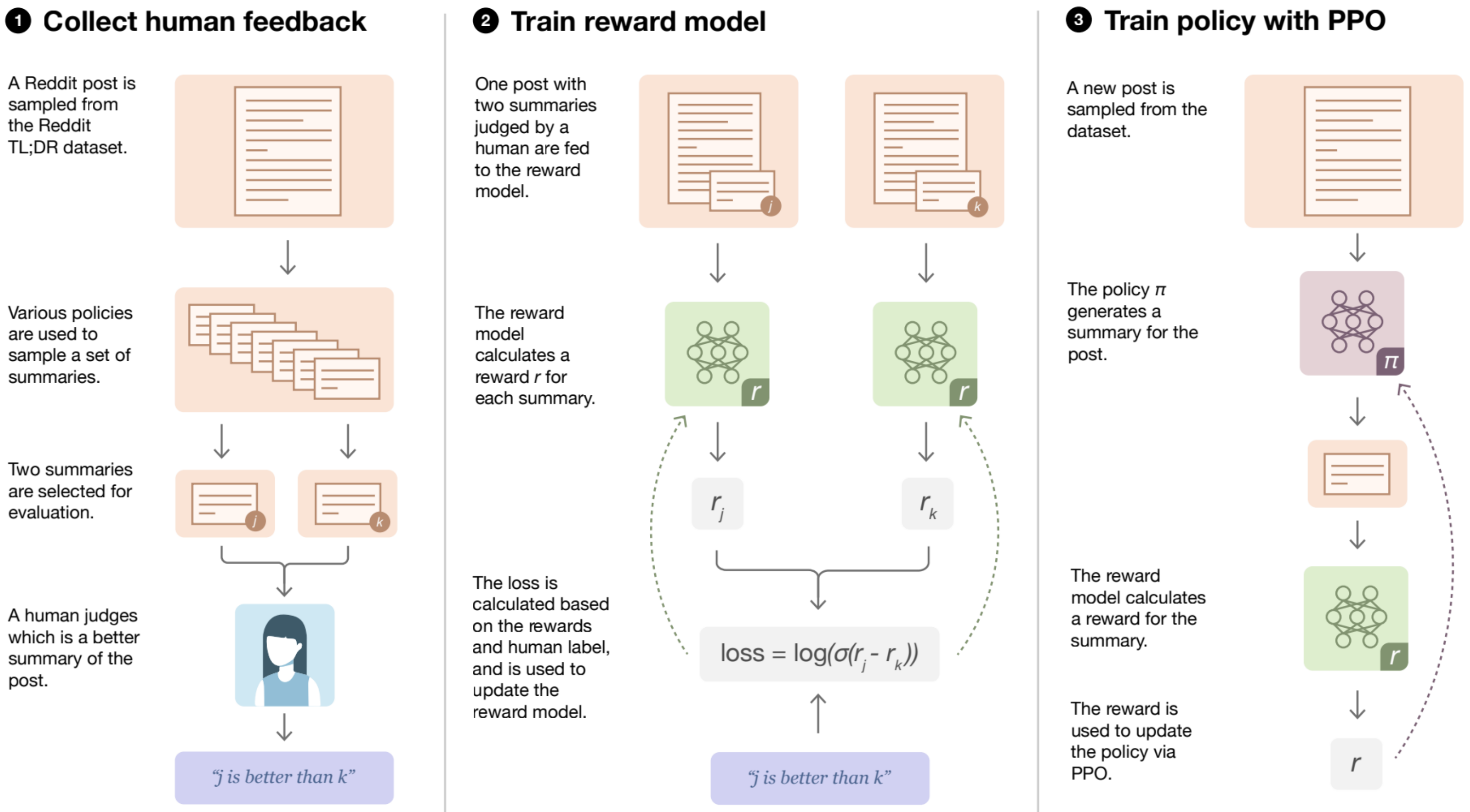


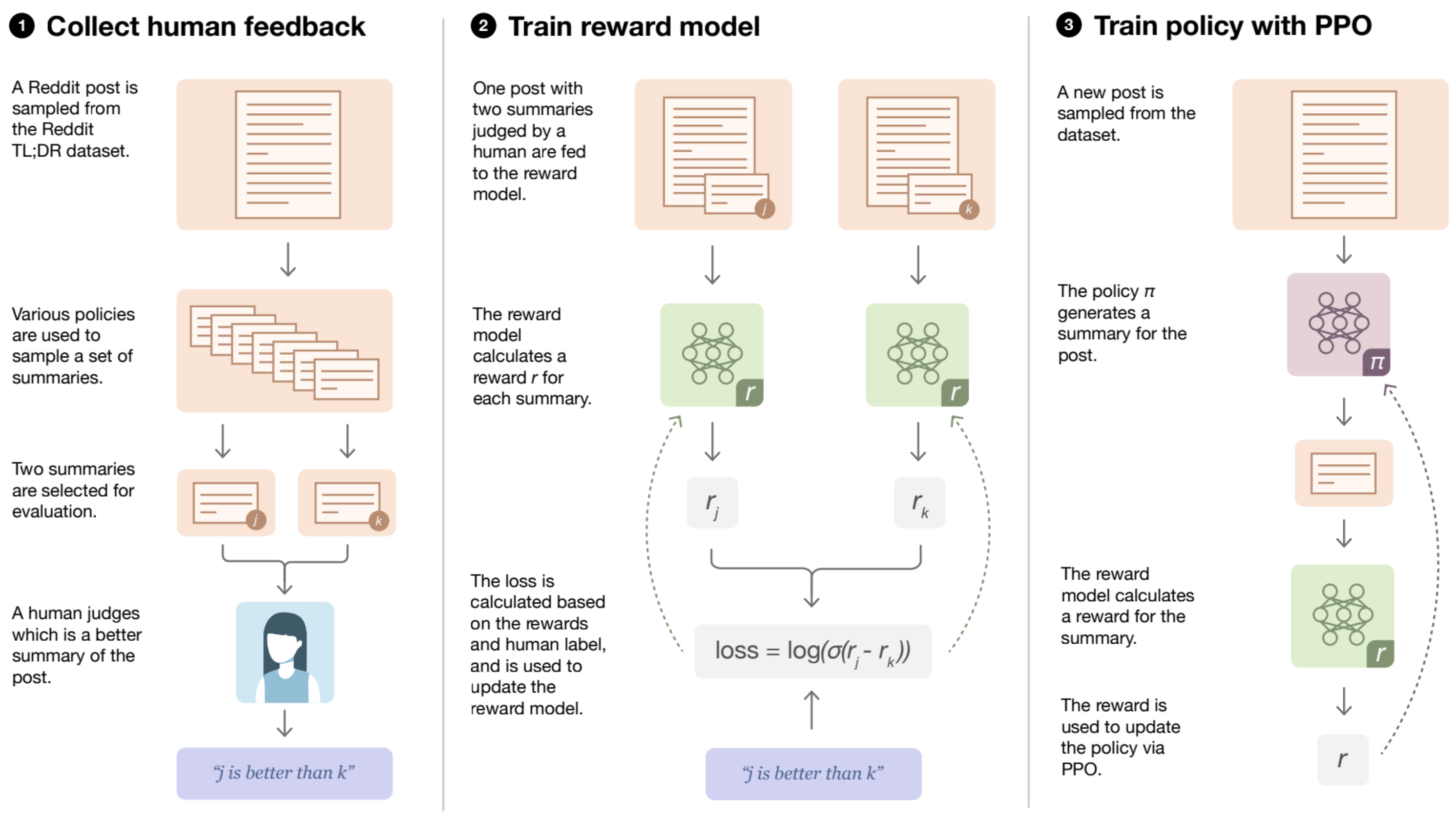
Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

Learning to summarize from human feedback(NeurIPS 2020), https://arxiv.org/pdf/2009.01325

RLHF(Reinforcement Learning from Human Feedback)

1) Supervised fine-tuning (SFT)

- fine-tuning a pre-trained LM with **supervised learning** on **high-quality data** for downstream task of interest(dialogue, summarization, etc.) to obtain a **model** $\boldsymbol{\pi^{SFT}}$

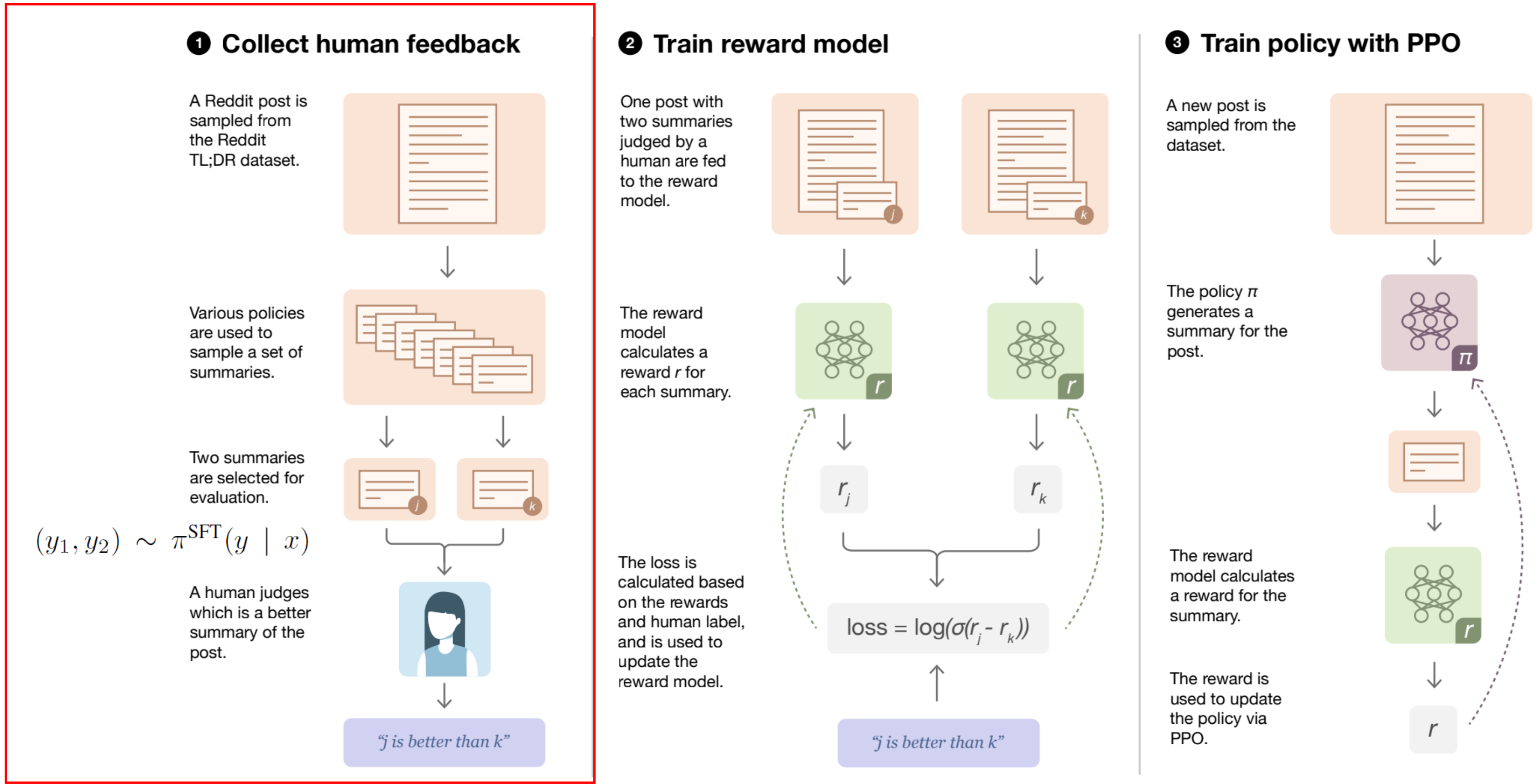


RLHF(Reinforcement Learning from Human Feedback)

2) Reward Modelling Phase

- SFT model is prompted with prompts x to produce pairs of answer $(y_1, y_2) \sim \pi^{\mathrm{SFT}}(y \mid x)$
- preferred completion: y_w, dispreferred completion: y_l $y_w \succ y_l \mid x$

$$\mathcal{D} = \left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$$



RLHF(Reinforcement Learning from Human Feedback)

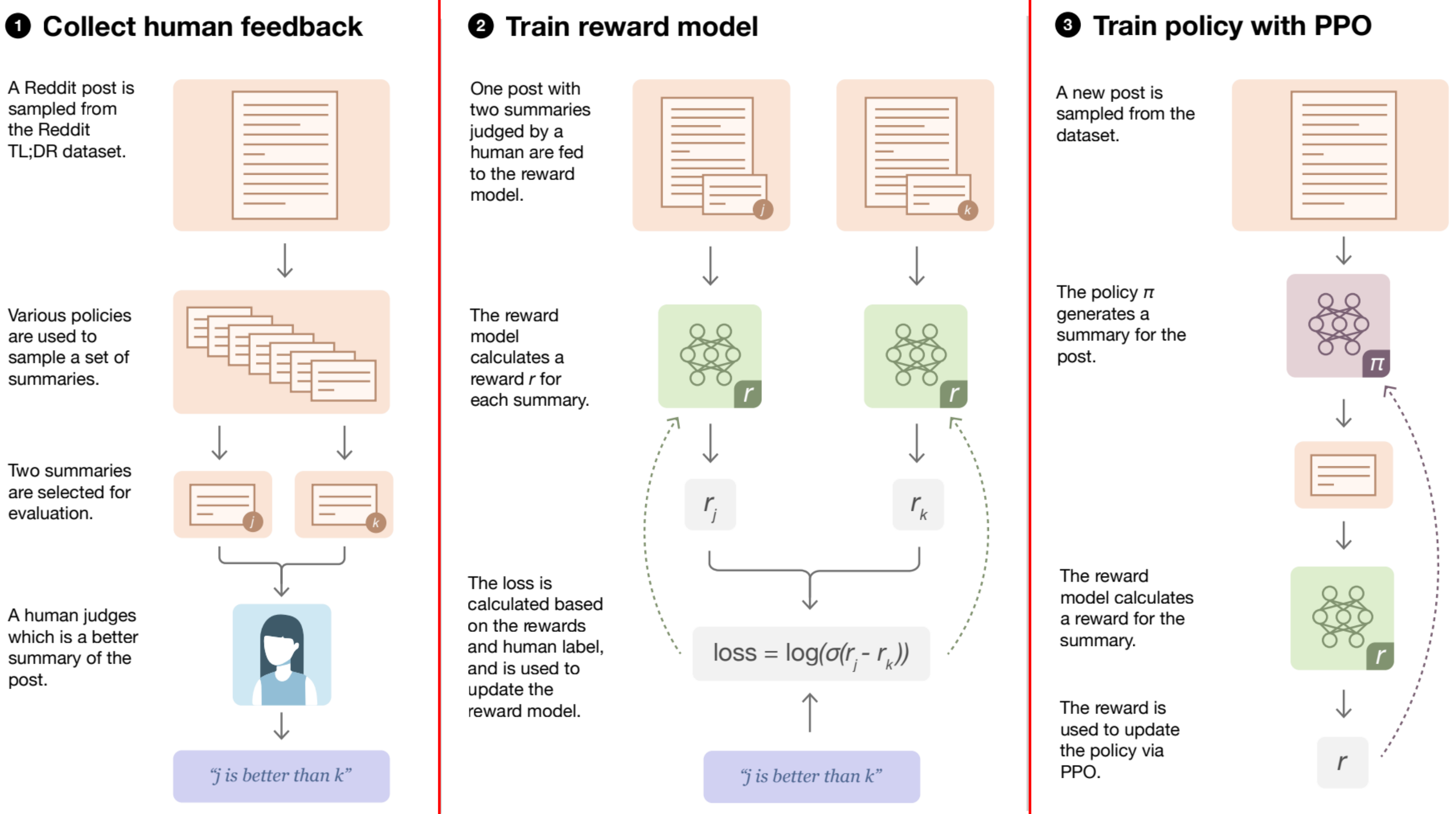
2) Reward Modelling Phase

- negative log-likelihood loss using Bradley-Terry

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

$$\mathcal{D} = \left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N} \qquad \mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)\right]$$

where σ is the logistic function.



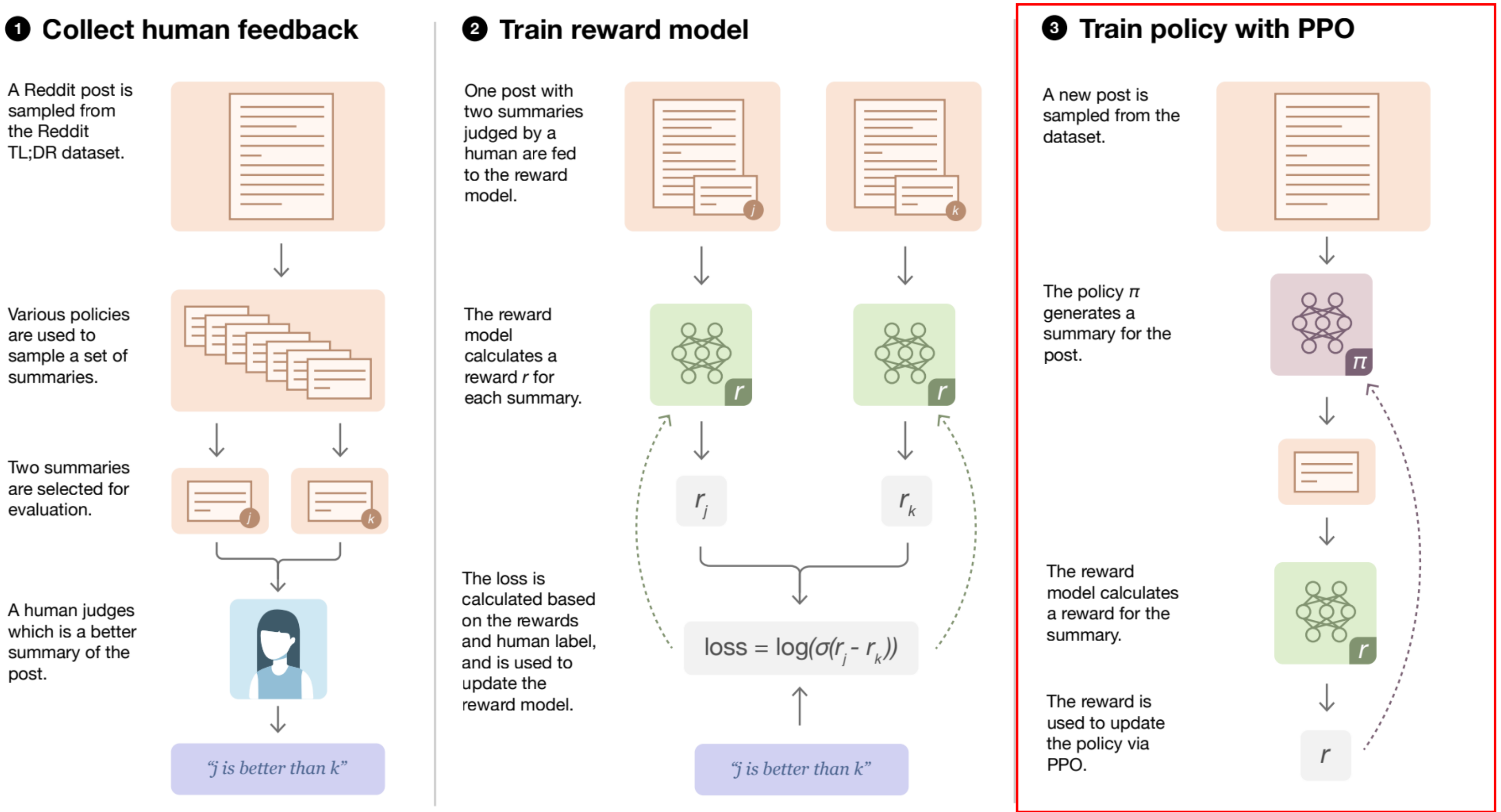
RLHF(Reinforcement Learning from Human Feedback)

3) RL Fine-Tuning Phase

- the learned reward function is used to provide feedback to the LM

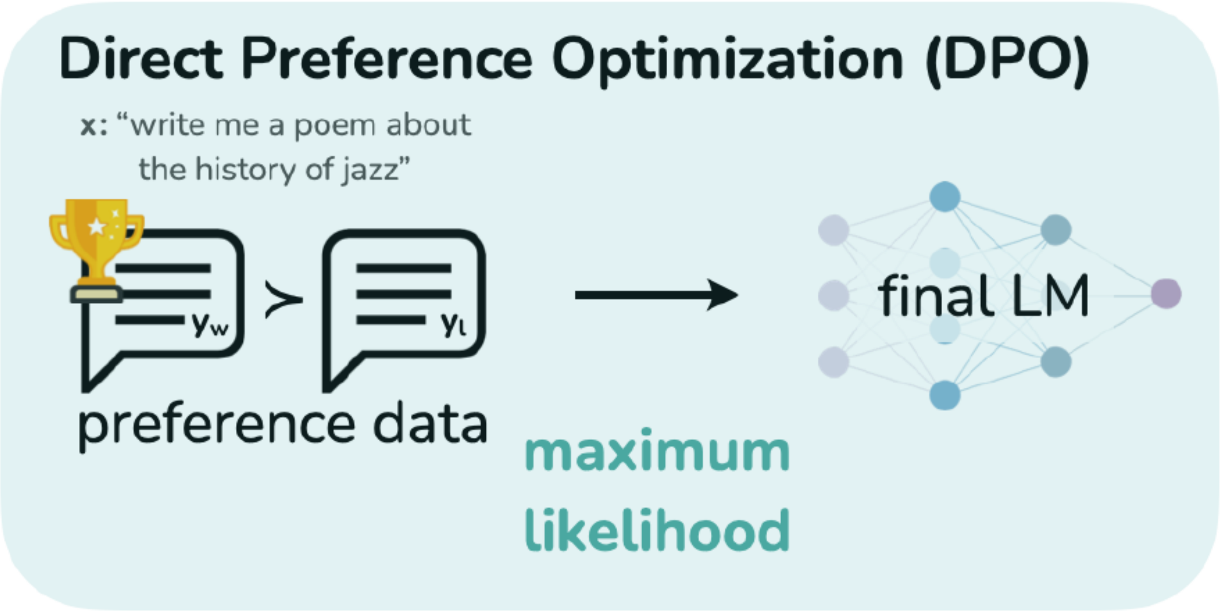
controlling the deviation from based reference policy $\pi_{ref} = \pi^{SFT}$

$$\max_{\pi_\theta} \underbrace{\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right]}_{\text{maximize reward}} - \beta \underbrace{\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]}_{\text{prevent the model from changing too drastically}},$$



DPO(Direct Preference Optimization)

- Leverages a particular choice of reward model parameterization that enables extraction of its optimal policy in closed form, without an RL training loop.
- Policy network represents both the LM and the (implicit) reward



DPO(Direct Preference Optimization)

RL objective under general reward function r

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

$$
\begin{aligned}
&= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x, y) - \beta\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \\
&= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta}r(x, y)\right] \\
&= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x, y)\right)} - \log Z(x)\right]
\end{aligned}
$$

$Z(x) = \sum_y \pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x, y)\right)$ is the partition function

DPO(Direct Preference Optimization)

optimal policy π_r using reward function r: KL-constrained reward maximization objective

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

reference policy π_{ref}

$Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$ is the partition function

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

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

logarithm of both sides → rearrange the optimal solution to express r(x,y)

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

DPO(Direct Preference Optimization)

logarithm of both sides → rearrange the optimal solution to express r(x,y)

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

The optimal RLHF policy π^* under BT model that satisfies the preference model:
reparameterization to ground-truth reward r^*, and optimal model π^*

$$p^*(y_1 \succ y_2 \mid x)=\sigma(r^*(x, y_1)-r^*(x, y_2))$$

express the human preference probability in terms of optimal policy π^* and reference policy π_{ref},

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

DPO(Direct Preference Optimization)

express the human preference probability in terms of optimal policy π^* and reference policy π_{ref},

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

DPO objective:

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

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

DPO(Direct Preference Optimization)

gradient of the loss function $\mathcal{L}_{DPO}$

- Importantly, the examples are weighted by how much higher the implicit reward model 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
  → Empirically import!

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

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

$$\text{where } \hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

DPO(Direct Preference Optimization)

- **Unlikelihood baseline** @sentimental experiments
  : simply maximizing $logp(y_w|x)$, the log probability of the preferred response, while minimizing $logp(y_l|x)$, the log probability of the dispreferred response

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
SUBREDDIT: r/tifu TITLE: TIFU by accidently kicking an old woman POST: So this didn't happen today but actually about a year or two ago. I was at my granddads funeral so of course it was all very sad and full of lots of crying old people. After the ceremony everyone walks outside the building and onto the other side of the small road the hearses drive down. Now the road is important because obviously if there's a road, there's a curb onto the sidewalk, so most of us are on the other side of the road, besides a few older people walking a lot slower. As one of the old woman goes to walk up the curb [. . .] TL;DR:	when an old woman was tripping the when

Table 3: Unlikelihood samples from TL;DR prompts sampled at temperature 1.0. In general, we find unlikelihood fails to generate meaningful responses for more complex problems such as summarization and dialogue.

DPO(Direct Preference Optimization)

1. **Sample completions $\boldsymbol{y_1}, \boldsymbol{y_2} \sim \boldsymbol{\pi_{ref}}(\cdot \mid \boldsymbol{x})$ 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 LM $\boldsymbol{\pi_\theta}$ to minimize $\boldsymbol{\mathcal{L}_{DPO}}$** for the given π_{ref} and $\mathcal{D}$ and desired β.

Reuse preference datasets available!

- the preference datasets are sampled using π^{SFT}, $\pi_{ref} = \pi^{SFT}$

Experiments

- Baselines (<6B)
  : GPT-J, Pythia-2.8B, SFT, Preferred-FT, Unlikelihood, PPO, PPO-GT, Best-of-N

- Task 1: controlled sentiment generation
  - x : a prefix of a movie review from IMDb dataset
  - policy must generate y with positive sentiment
  - generate preference pairs using pre-trained sentiment classifier
  - SFT: fine-tune GPT-2-large

- Evaluation
  - controlled sentiment generation: using ground-truth reward function (the pre-trained sentiment classifier)
  - Real world: win rate against a baseline policy(using GPT-4)

49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his par...	negative

Experiments

- Task 2: Summarization
  - x : forum post from Reddit
  - policy must generate a summary y of the main point of the post
  - use the Reddit TL;DR summarization dataset
  - SFT model finetuned on human-written forum post summaries with TRLX for RLHF

- Evaluation
  - use references summaries in the test set

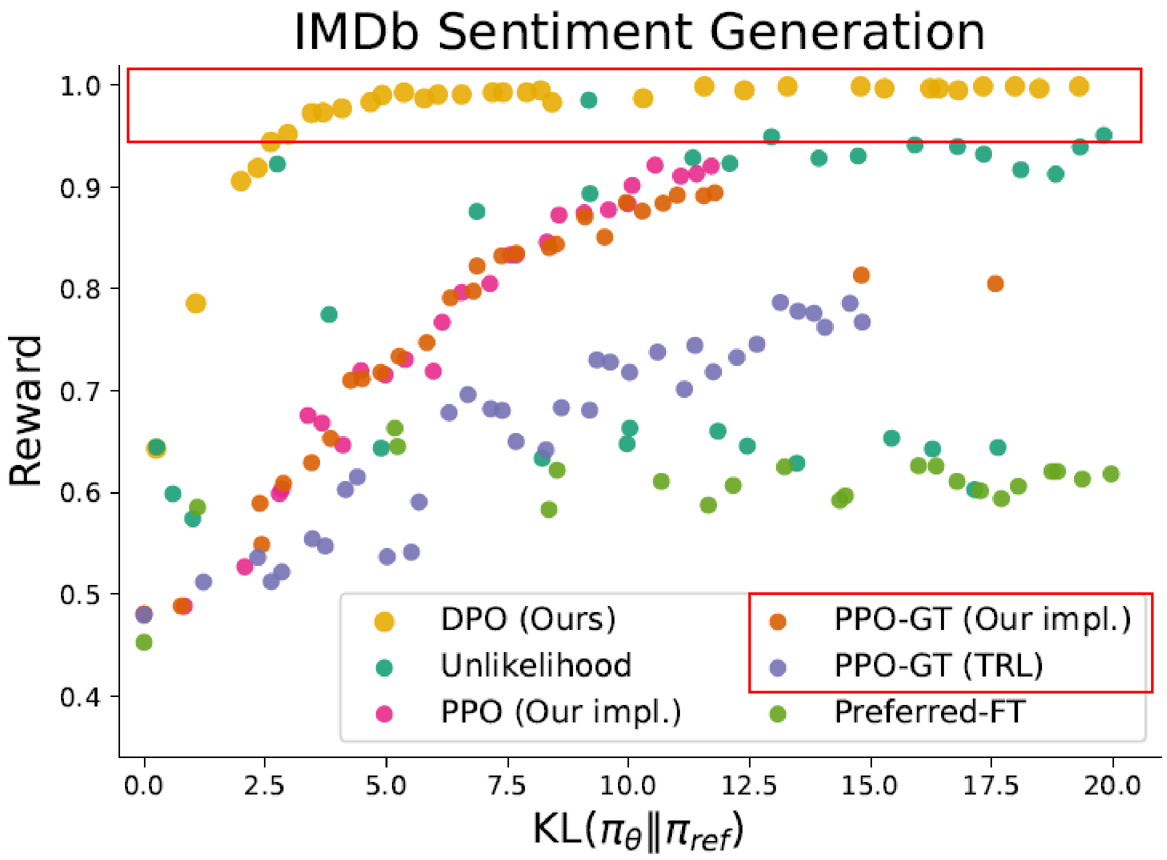
Experiments

- Task 3: Single-turn dialogue
  - x : human query, which may be anything
    (from a question about astrophysics to a request for relationship advice)
  - policy must generate an engaging and helpful response y
  - Anthropic Helpful and Harmless dialogue dataset (170K)
  - No pretrained SFT model is available

- Evaluation
  - use preferred response in the test dataset

Experiments

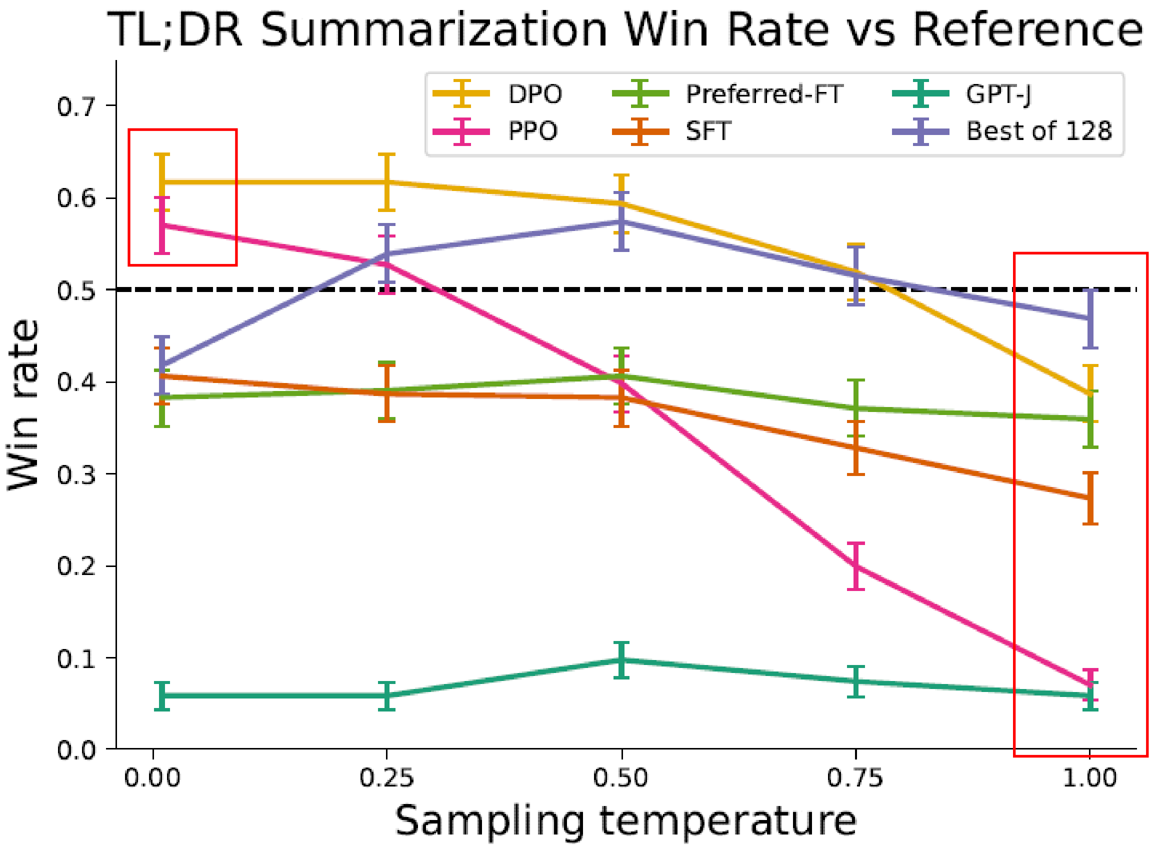
1) How well can DPO optimize the RLHF objective?



Experiments

2) Can DPO scale to real preference datasets ? (summarization and single-turn dialog)

DPO win rate 61% at temperature of 0



DPO, PPO, Preferred-FT
- based on GPT-J SFT model

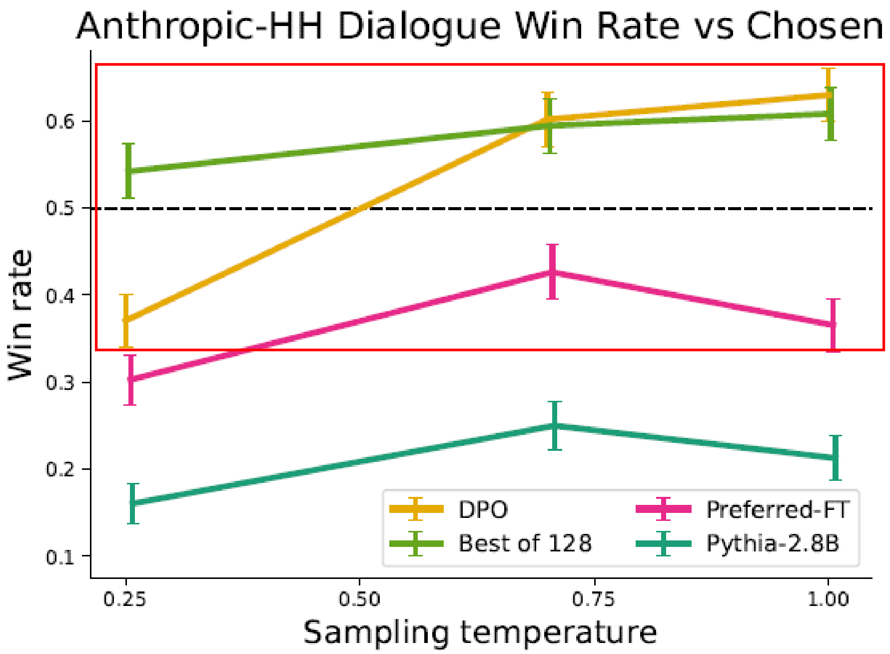
temperature
- 0: deterministic
- 1: dynamic

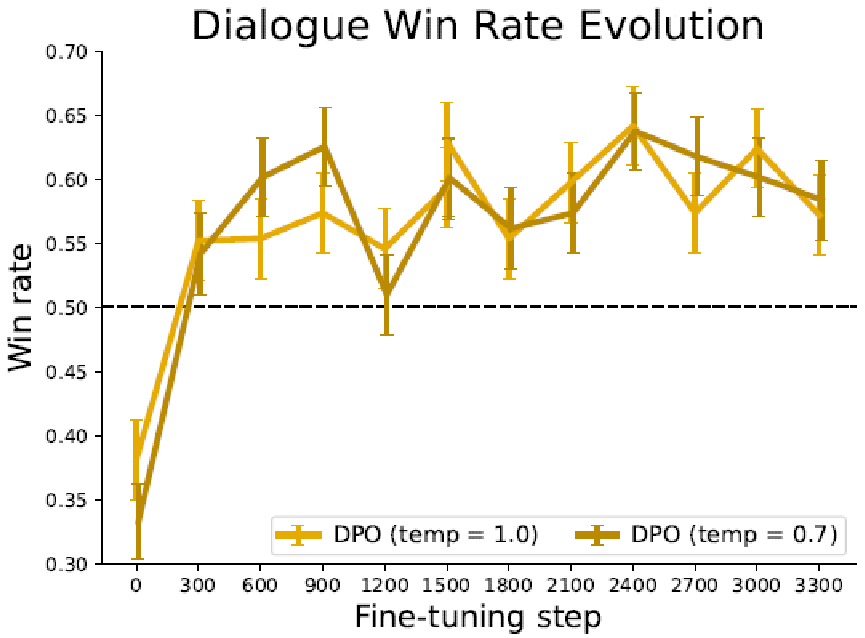
Experiments

2) Can DPO scale to real preference datasets ? (summarization and single-turn dialog)

- no standard SFT model → pre-train Pythia-2.8B, Preferred-FT to train a reference model

DPO is the only **computationally efficient** method that improves over the preferred completions





RLHF model trained with **PPO is unable to find** a prompt or sampling temperature that gives performance better than the base Pythia-2.8B model

Experiments

3) Generalization to a new input distribution

- the PPO and DPO policies from **Reddit TL;DR summarization** experiment on a distribution, new articles in the test split of the **CNN/DailyMail dataset**

	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.

DPO(Direct Preference Optimization)

- a simple training paradigm **for training language models from preferences without RL**.
- DPO maps between language model policies and reward functions that enables training a LM to satisfy human preference directly, with simple cross-entropy loss, without RL.

- Limitations & Future Work
  - How does the DPO policy generalization out of distribution, compared with an explicit reward function?
  - How does reward over-optimization manifest in the DPO setting?
  - Need to explore scaling DPO to state-of-the-art models larger than 6B
  - Need to study best way to elicit high-quality judgments (e.g. GPT-4)