# Theoretical Understanding of Learning from Human Preferences

M. G. Zar, et. al. R. Munos

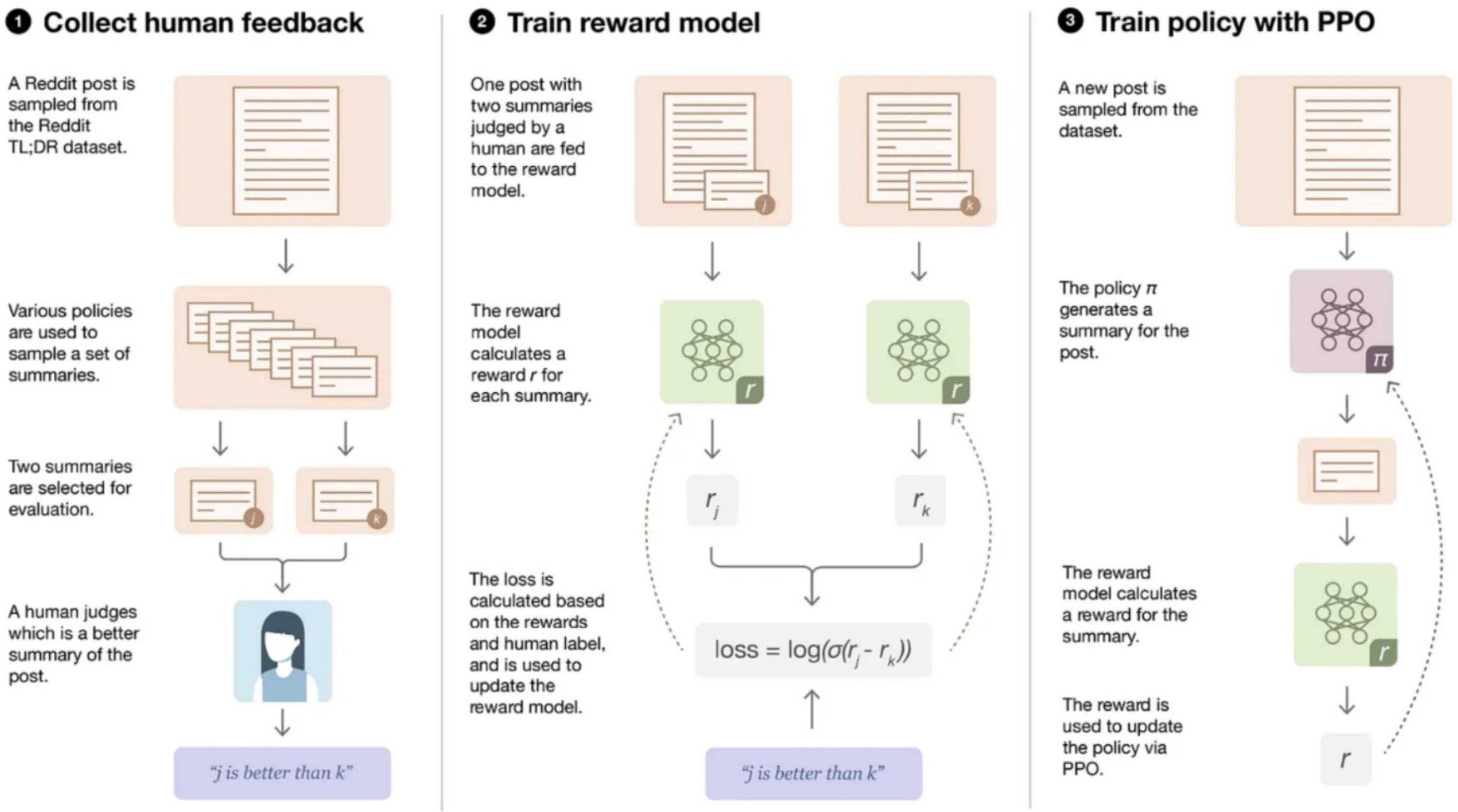
Deepmind

# Outline

- Background
  - RLHF PPO
  - DPO

- $\Psi$ Preference Optimization
  - Human Preference
  - $\Psi$PO
  - Special case: RLHF and DPO
  - Special case: IPO

# RLHF PPO


1 Collect human feedback
A Reddit post is sampled from the Reddit TL;DR dataset.
Various policies are used to sample a set of summaries.
Two summaries are selected for evaluation.
A human judges which is a better summary of the post.
"j is better than k"
2 Train reward model
One post with two summaries judged by a human are fed to the reward model.
The reward model calculates a reward r for each summary.
The loss is calculated based on the rewards and human label, and is used to update the reward model.
r
r
$r_j$
$r_k$
$loss = log(\sigma(r_j - r_k))$
"j is better than k"
3 Train policy with PPO
A new post is sampled from the dataset.
The policy π generates a summary for the post.
π
The reward model calculates a reward for the summary.
r
The reward is used to update the policy via PPO.
r

# Direct Preference Optimization (DPO)

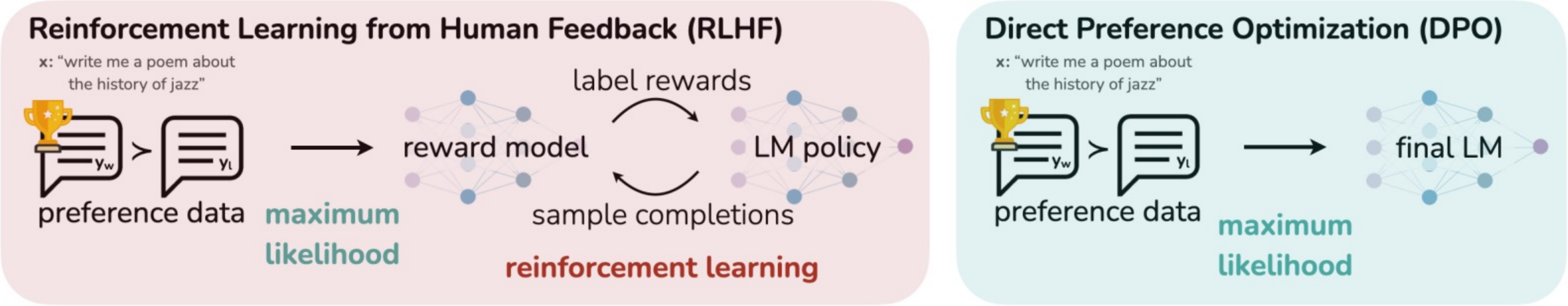



Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, without an explicit reward function or RL.

# Motivations : Two Assumptions ?

***Can we learn from human preference without two following assumptions?***

- 1. Pairwise preference using a **proxy pointwise reward**
  - not direct use human preference
  - e.g. elo score as reward in DPO, PPO

- 2. **Reward model Learning**
  - We need to train a reward model based on preference dataset, hope to be able to generalize to out of distribution data
  - E.g. Bradley Terry model in RLHF PPO

# Human Preferences

- The probability of human preferences of pairwise responses

$$p^*(y > y'|x) = \mathbb{E}_h\left[1(\text{h preferences y to y' given x})\right]$$

- Sample data generated by human preferences with probability

$$\mathcal{D} = \left(x_i, y_i, y_i', 1(y_i > y_i')\right)_{i=1}^{N} = (x_i, y_{i,w} > y_{i,l})_{i=1}^{N}$$

- Total preference of two policies

$$p_\rho^*(\pi > \mu) = \mathbb{E}_{x\sim\rho, y\sim\pi, y'\sim\mu)}\left[p^*(y > y'|x))\right]$$

# $\Psi$ Preference Optimization ($\Psi$PO)

- General RLHF Objective function
  - Use non-decreasing, non-linear function of human preference
  - Maximize of human preference
  - KL regularization, encouraging close to reference policy

$$\max_{\pi} \mathbb{E}_{x\sim\rho, y\sim\pi, y'\sim\mu)} \left[\Psi(p^*(y > y'|x))\right] - \tau D_{KL}(\pi||\pi_{ref})$$

# $\Psi$PO Special case: RLHF PPO

- Learning **Bradley Terry Reward model** for Pairwise Preference
- Use **rewards** of pairwise preferences as **proxy** of human preference

$$\mathcal{L}(r) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log p(y_w > y_l|x)\right]$$
$$p(y > y'|x) = \sigma(r(x,y) - r(x,y'))$$

- Policy Optimization with PPO

$$\mathcal{J}(\pi) = \mathbb{E}_{\pi}\left[r(x,y)\right] - \tau D_{KL}(\pi||\pi_{ref})$$

# $\Psi$PO Special case: DPO

- Bypass the stage of reward model learning
- But use **relative rewards** of pairwise preferences as **proxy** of human preference

$$p(y > y'|x) = \sigma(r(x,y) - r(x,y'))$$
$$\Psi(p) = \log p/(1-p)$$

- The Cross-Entropy Loss

$$\min_{\pi} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log\sigma(\tau\log\left(\pi(y_w|x)/\pi(y_l|x)\right) - \tau\log\left(\pi_{ref}(y_w|x)/\pi_{ref}(y_l|x)\right)\right]$$

# Identity Preference Optimization (IPO)

- Direct use **human preference** function
- **No reward model** learning
- Let the non-decreasing function $\Psi$ to be Identity function

$$\max_{\pi} p_{\rho}^{*}(\pi > \mu) - \tau D_{KL}(\pi || \pi_{ref})$$

# Sampled IPO Algorithm

---

**Algorithm 1** Sampled `IPO`

---

**Require:** Dataset $\mathcal{D}$ of prompts, preferred and dispreferred generations $x$, $y_w$ and $y_l$, respectively. A reference policy $\pi_{\text{ref}}$

1: Define

$$h_\pi(y, y', x) = \log\left(\frac{\pi(y|x)\pi_{\text{ref}}(y'|x)}{\pi(y'|x)\pi_{\text{ref}}(y|x)}\right)$$

2: Starting from $\pi = \pi_{\text{ref}}$ minimize

$$\mathop{\mathbb{E}}_{(y_w, y_l, x) \sim D}\left(h_\pi(y_w, y_l, x) - \frac{\tau^{-1}}{2}\right)^2.$$

# Weak Regularization and Overfitting

- DPO prune to overfitting
  - The strength of the KL-regularization becomes weaker and weaker, the more deterministic of the preferences

$$p(y > y'|x) \to 1, r(x, y) - r(x, y') \to \infty$$

$$\pi(y'|x) = 0$$

- IPO to remedy
  - Bound the human reference function $\Psi$
  - Ensure the KL regularization remains effective
  - Regressing the gap between log-likelihood ratio to regularization strength

$$\log(\pi(y_w)/\pi(y_l)) - \log(\pi_{ref}(y_w)/\pi_{ref}(y_l)) \to \tau^{-1}/2$$

# References


- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, Rémi Munos **A General Theoretical Paradigm to Understand Learning from Human Preferences**, 2023
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, **Direct Preference Optimization: Your Language Model is Secretly a Reward Model**, NeurIPS 2023
- L. Ouyang, et. al OpenAI, **Training language models to follow instructions with human feedback**, NeurIPS 2022