Group Robust Preference Optimization in Reward-free RLHF

Shyam S. Ramesh

Yifan Hu

Iason Chaimalas

Viraj Mehta

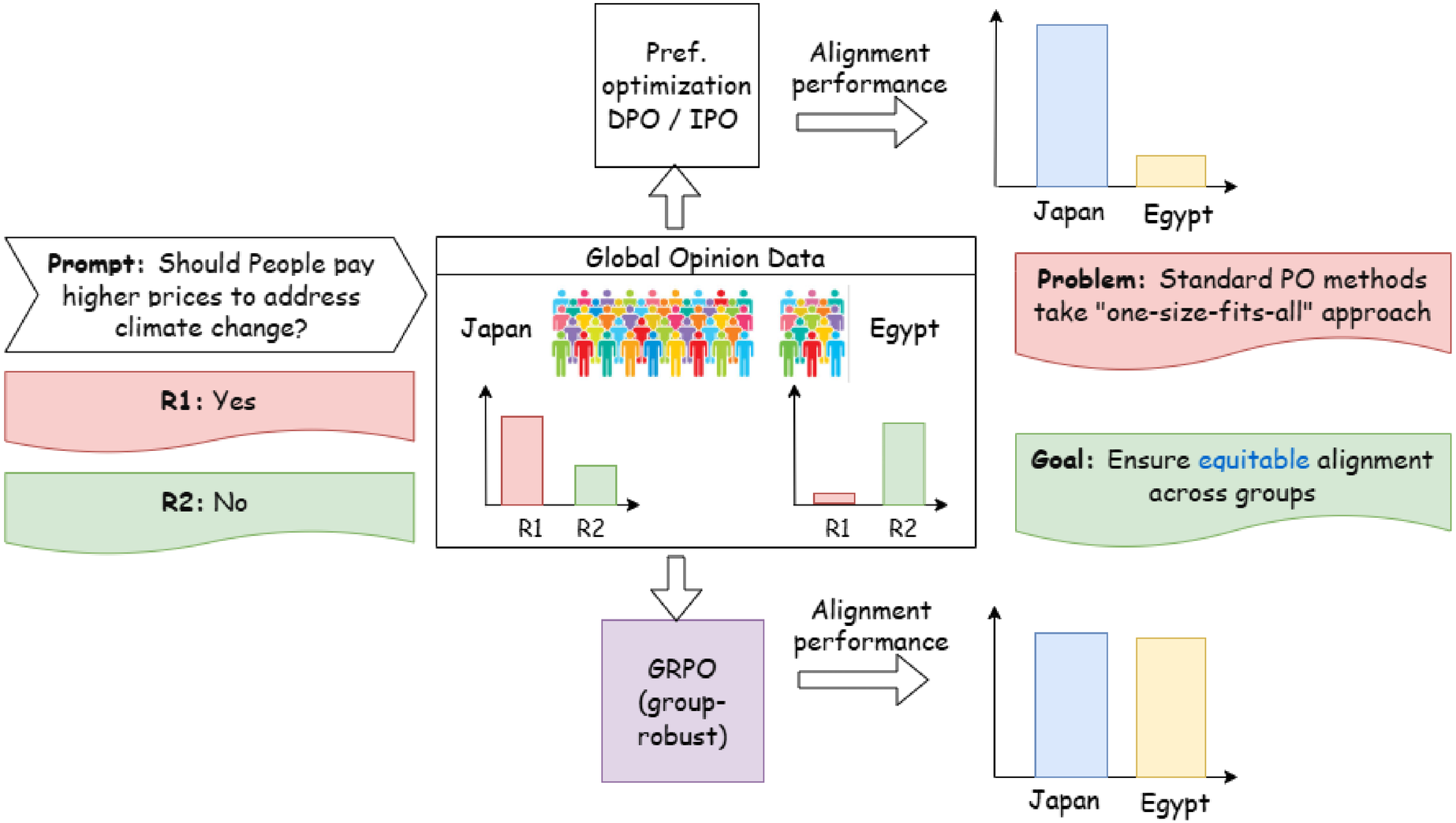
Pier Giuseppe Sessa

Haitham Bou Ammar

Ilija Bogunovic

Motivation

Fine-tuning a large language model (LLM) to align with diverse user preferences



Tackling challenges

- Challenge 1: Preference data come from diverse groups
  - ☑ Include group information in the context of the LLM

Preference data with group info.: $(x, y_w, y_l), \ g \in \mathcal{G}$

Latent reward: $\mathrm{r}(\mathrm{x_g}, \mathrm{y}) \ \text{ where } \ x_g = x \oplus g$

Preference probability derived from the BT model: $p(y > y' | x_g) = \frac{\exp(r(x_g, y))}{\exp(r(x_g, y)) + exp(r(x_g, y'))}$

- Challenge 2: Uneven distribution/quantity among groups
  - ☑ Optimize against worst-case alignment performance across *all* groups

Robust loss for reward learning: $\mathcal{L}_R = \max_{g \in \mathcal{G}} -\mathbb{E}_{(x_g, y_w, y_l) \in \mathcal{D}_g} [\log \sigma(r(x_g, y_w) - r(x_g, y_l))]$

Group robust direct preference optimization

DPO

- KL-regularized reward optimization objective for policy

$$\pi^{\star} = \arg\max_{\pi}[\mathbb{E}_{x,y\sim\pi(\cdot|x)}[r(x,y)] - \beta D_{KL}[\pi(y|x)||\pi_{ref}(y|x)]]$$

- Reward in terms of optimal policy (DPO - Rafailov *et al.*'23):

$$r(x,y) = \beta \log\left(\frac{\pi^{\star}(y|x)}{\pi_{ref}(y|x)}\right) + \beta \log Z(x)$$

Substitute $r(x,y)$ in robust reward loss to obtain the group-robust DPO loss

$$\mathcal{L}_{GRPO} = \max_{g\in\mathcal{G}} \mathbb{E}_{(x_g,y_w,y_l)\sim\mathcal{D}_g} - \log\sigma\left(\beta \log\frac{\pi(y_w|x_g)}{\pi_{ref}(y_w|x_g)} - \beta \log\frac{\pi(y_l|x_g)}{\pi_{ref}(y_l|x_g)}\right)$$

(DPO loss can be replaced with IPO loss, hinge loss, etc. (Tang et al., 2024))

Algorithm outline (GR-DPO)

Min-max game
(best policy for worst group alignment)

Per-group weights

Group conditioned language policy

$$\min_{\pi_\theta} \max_{\alpha \in \Delta^{K-1}} \sum_{g=1}^{K} -\alpha_g \mathbb{E}_{(x_g, y_w, y_l) \sim \mathcal{D}_g} \log \sigma \left(\beta \log \frac{\pi(y_w|x_g)}{\pi_{ref}(y_w|x_g)} - \beta \log \frac{\pi(y_l|x_g)}{\pi_{ref}(y_l|x_g)} \right)$$

Per-group DPO Loss

Group weights
over a simplex

Language policy

Algorithm outline (GR-DPO)

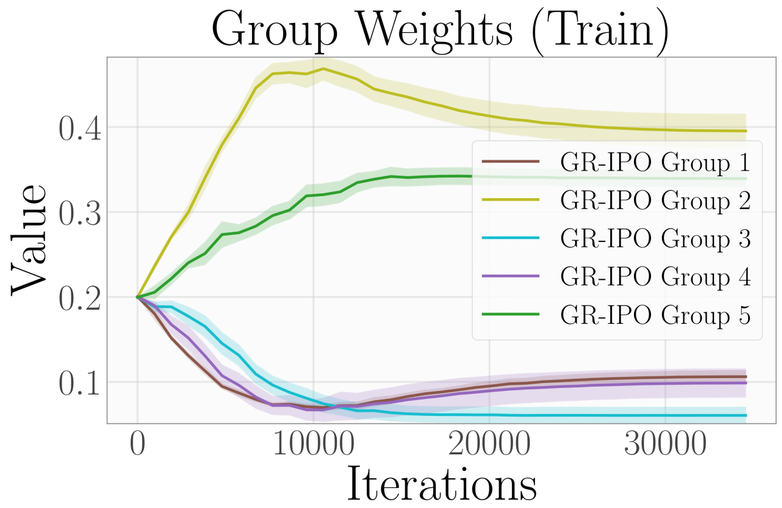
$$\min_{\pi_\theta} \max_{\alpha \in \Delta^{K-1}} \sum_{g=1}^{K} -\alpha_g \mathbb{E}_{(x_g, y_w, y_l) \sim \mathcal{D}_g} \log \sigma \left(\beta \log \frac{\pi(y_w|x_g)}{\pi_{ref}(y_w|x_g)} - \beta \log \frac{\pi(y_l|x_g)}{\pi_{ref}(y_l|x_g)} \right)$$

Alternating algorithm:

1)Gradient descent update on policy parameters θ

2)Multiplicate weights update for α

Groups with higher cumulative loss have higher α_g

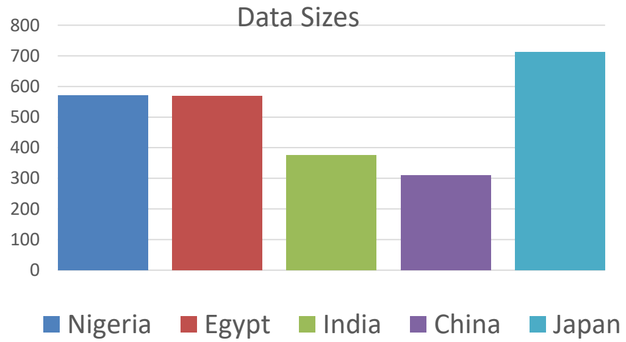


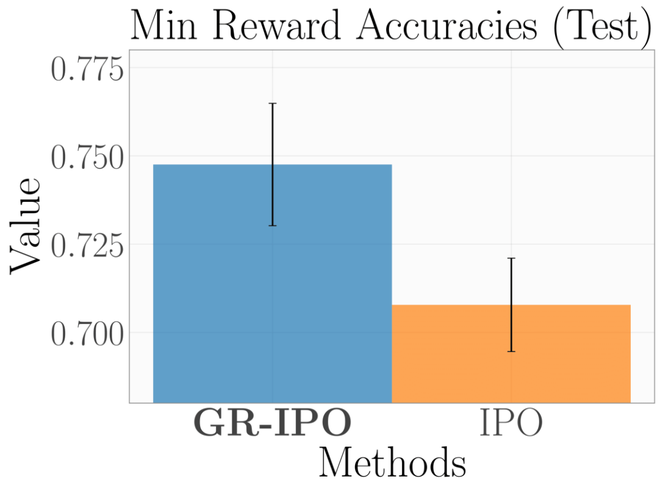
Convergence guarantees $\mathcal{O}(T^{-1/2})$ in case of log-linear policies $\pi_\theta(a|x) = \frac{exp(\theta^T \phi(x,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^T \phi(x,a))}$

Experiments

GlobalOpinionQA dataset with Gemma-2B model

- 5 Groups = Countries
- Align LLM to each country preference robustly





Our GR-IPO improves loss and boosts reward accuracy of worst group

Take-home Messages

- The first study on group robustness in RLHF
- To robustly align a LLM to diverse user groups
  - No need for novel PO losses!
  - Simply using carefully weighted DPO and IPO suffices

Group Robust Preference Optimization (GRPO)

Blog

Check out our poster on Fri 13 Dec, 1 - 4 p.m.

Read our paper on arxiv (QR link)

Read our blog (QR link)

Reach out to shyam.ramesh.22@ucl.ac.uk

Paper