ICLR

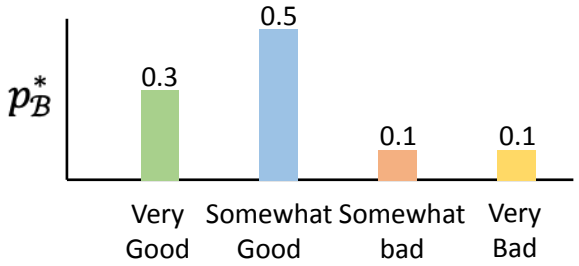
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

No Preference Left Behind: **Group Distributional Preference Optimization**

Binwei Yao, Zefan Cai, Sean Yun-Shiuan Chuang, Shanglin Yang,
Ming Jiang, Diyi Yang, Junjie Hu

Motivation

x: In your opinion, what is the significance of the availability of products from different parts of the world for our country?



Distribution of Opinions

$b_c \sim p^*_{\mathcal{B}}$ | $b_r \sim \mathcal{B} \setminus \{b_c\}$

Somewhat Good | *Somewhat Bad*

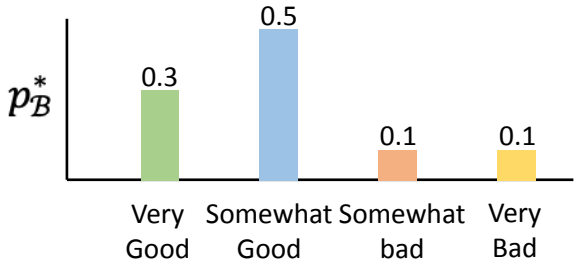
y_c > y_r

I believe that, in many circumstances, the access to a wide variety of products from different parts of the world is a sign of growth and evolution.

It definitely has its pros and cons. It's nice to have access to all these products, but we need to make sure we're supporting our local economy as well.

Motivation

x: In your opinion, what is the significance of the availability of products from different parts of the world for our country?



Distribution of Opinions

$b_c \sim p^*_{\mathcal{B}}$ | $b_r \sim \mathcal{B}\backslash\{b_c\}$

Somewhat Good | *Somewhat Bad*

y_c > y_r

I believe that, in many circumstances, the access to a wide variety of products from different parts of the world is a sign of growth and evolution.

It definitely has its pros and cons. It's nice to have access to all these products, but we need to make sure we're supporting our local economy as well.

What if $y_r > y_c$ co-exist?

Why not DPO?

Direct Preference Optimization (DPO)

$$\ell_{\text{DPO}} = \ell_{\text{pref.}}(y_c > y_r, x)$$

Preference Alignment

Instance-level Preferences

$P(b_i) = 0.5$ $\quad p_\theta(y_i|x) > p_\theta(y_j|x)$

$P(b_j) = 0.1$ $\quad p_\theta(y_j|x) > p_\theta(y_i|x)$

Conflicting Preferences

Loss Function of DPO

$$\ell_{\text{dpo}}(y_c \succ y_r, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)}\right)\right]$$

Why not DPO?

Loss Function of DPO

$$\ell_{\text{dpo}}(y_c \succ y_r, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \right) \right]$$

Chosen Response

Rejected Response

Why not DPO?

Loss Function of DPO

$$\ell_{\text{dpo}}(y_c \succ y_r, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \right) \right]$$

Conflicting Preferences

$$\ell_{\text{dpo}}(y_r \succ y_c, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} - \beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} \right) \right]$$

Why not DPO?

Loss Function of DPO

$$\ell_{\text{dpo}}(y_c \succ y_r, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \right) \right]$$

Conflicting Preferences

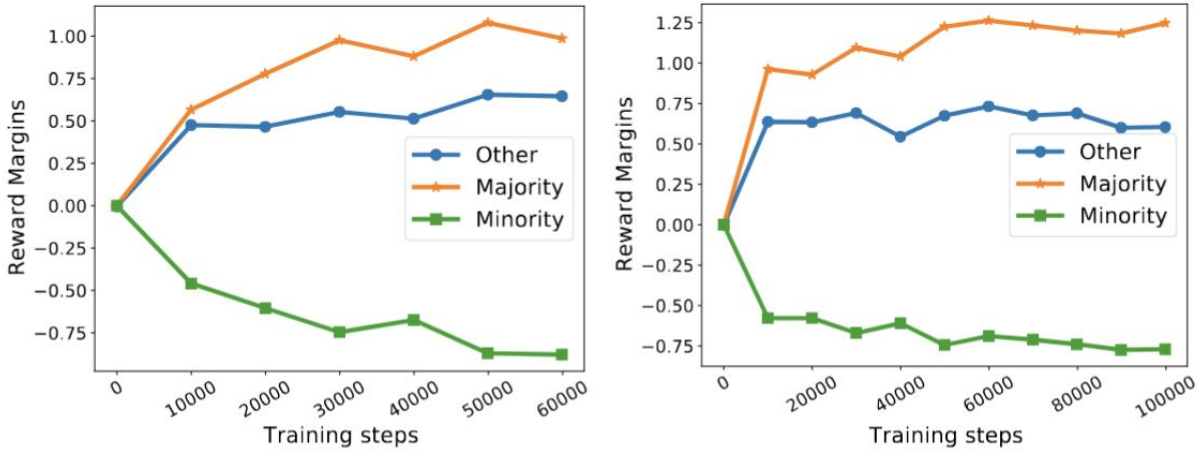
$$\ell_{\text{dpo}}(y_r \succ y_c, x; \theta) = -\mathbb{E}_{(x,y_c,y_r)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} - \beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} \right) \right]$$

Conflicting preferences can cancel each other out

Why not DPO?

Reward Margins = Chosen Rewards - Reject Rewards

$$R(x, y_c, y_r) = r(x, y_c) - r(x, y_r) = \beta \log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)}$$



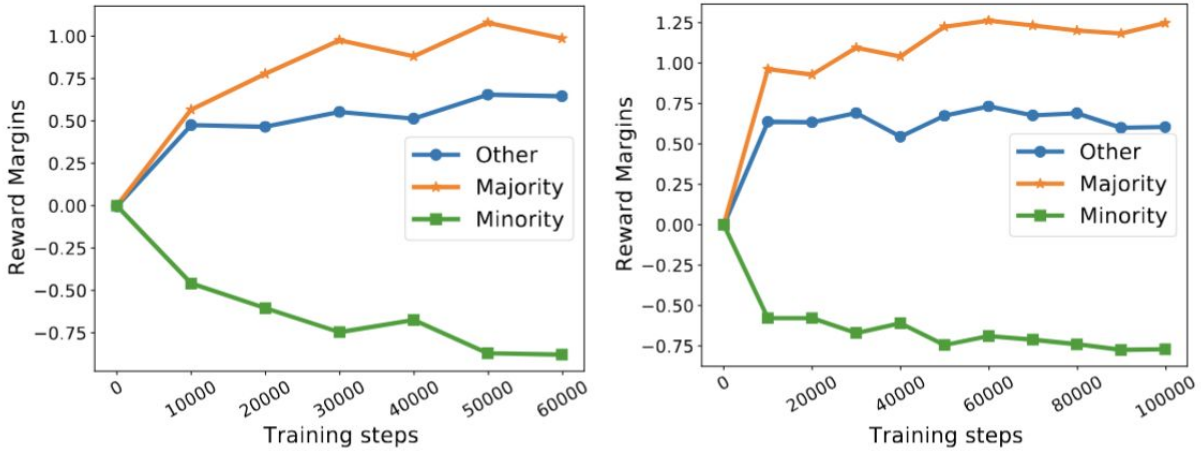
(a) GPT-2 Large

(b) Pythia-2.8B

Why not DPO?

Reward Margins = Chosen Rewards - Reject Rewards

DPO Skews Towards Majority Preferences !!!



(a) GPT-2 Large

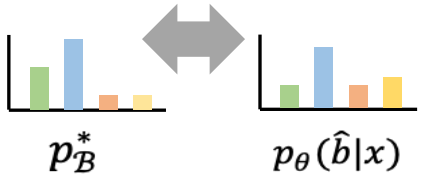
(b) Pythia-2.8B

GDPO: from *Instance-level* to *Distributional-level* Alignment

Group Distributional Preference Optimization (GDPO)

$$\ell_{\text{GDPO}} = \ell_{\text{cal.}}(p_\theta(\hat{b}|x), p^*_{\mathcal{B}}) + \ell_{\text{pref.}}(y_c > y_r, b_c, x)$$

KL Divergence



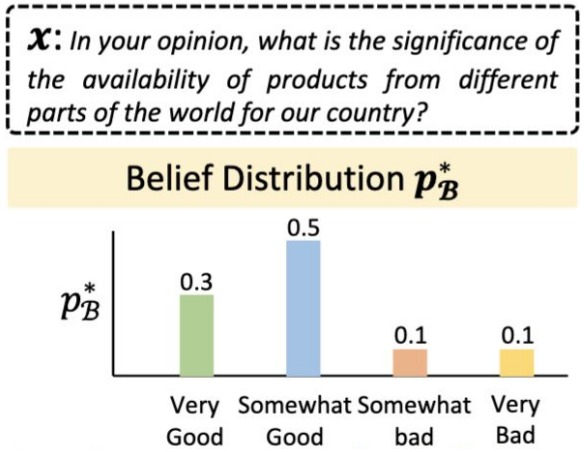
Belief-Conditioned Preference Alignment

$$p_\theta(y_i|x, b_i) > p_\theta(y_j|x, b_i)$$

Group Distributional Preference Optimization

What's belief?

- **Belief as the degree to which individuals agree with a particular stance.** While individual preferences may vary depending on the context, beliefs can be represented as the extent of agreement with the statements in preference-related sentences.



Group Distributional Preference Optimization

GDPO Training Objective: Two-step Prediction

$$p_\theta(y|x) = \sum_{b \in \mathcal{B}} p_\theta(y|b,x) p_\theta(b|x)$$

GDPO Loss

$$\ell_{\text{gdpo}}(x, p^*_{\mathcal{B}}, y_{\mathcal{B}}; \theta) = \underbrace{\ell_{\text{cal.}}(p_\theta(b|x), p^*_{\mathcal{B}})}_{\text{belief calibration loss}} + \underbrace{\mathbb{E}_{b_c \sim \mathcal{B}, y_c, y_r \sim y_{\mathcal{B}}} \ell_{\text{pref}}(y_c \succ y_r, b_c, x)}_{\text{belief-conditioned preference alignment loss}},$$

Group Distributional Preference Optimization

Calibration Loss

$$\ell_{\text{cal.}}(p_\theta(b|x), p^*_{\mathcal{B}}) = \mathbf{KL}(p_\theta(b_{[0]}|x), p^*_{\mathcal{B}}) - \log p_\theta(b|x)$$

Belief-conditioned Alignment Loss

$$\ell_{\text{pref}}(y_c \succ y_r, b_c, x) = -\log \sigma \left(\beta \log \frac{p_\theta(y_c \mid x, b_c)}{p_{\text{ref}}(y_c \mid x, b_c)} - \beta \log \frac{p_\theta(y_r \mid x, b_c)}{p_{\text{ref}}(y_r \mid x, b_c)} \right)$$

Experiment Setting

- **Controllable opinion generation**
  - Synthetic data generated from GlobalOpinionQA (Multi-Choice QA)
- **Controllable review generation**
  - Real-world data from Amazon Movie Review

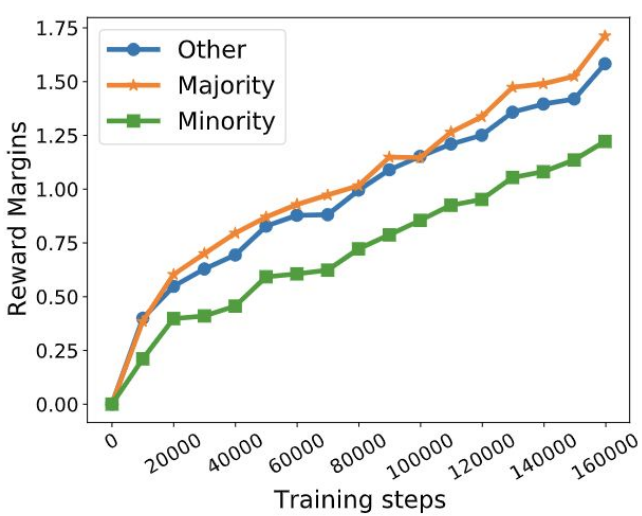
Split	Unite States (469)		Pakistan (219)		S.Africa (162)	
	small	large	small	large	small	large
Train	14,321	176,905	6,684	80,364	4,960	54,896
Eval	1,843	22,166	860	10,070	636	6,878
Test	1,843	22,199	876	10,086	648	6,890

Split	Movies (692)	
	Small	**Large**
Train	13,825	73,804
Eval	1,657	9,155
Test	2,406	10,114

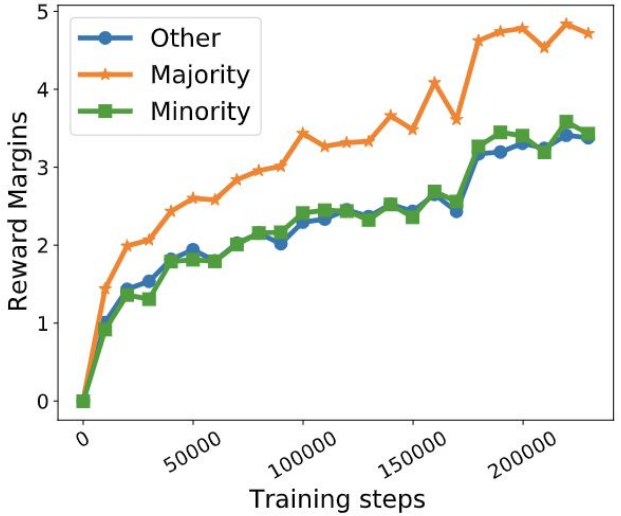
Evaluation Metrics

- **Belief Calibration Evaluation**
  - **JSD**: quantify the divergence between the ideal distribution and the learned distribution for each question.
  - **CBC**: measures the consistency between the predicted belief class token and the predicted belief description.
- **Conditioned Preference Generation Evaluation**
  - **BPC:** measures the consistency between the predicted belief and the predicted preferred response
  - **RS**: evaluate the cosine similarity scores between the embeddings of the model output and the human-written reference that expresses the same opinion as the model predicts.

Finding 1: GDPO Optimizes Minority Preference Learning

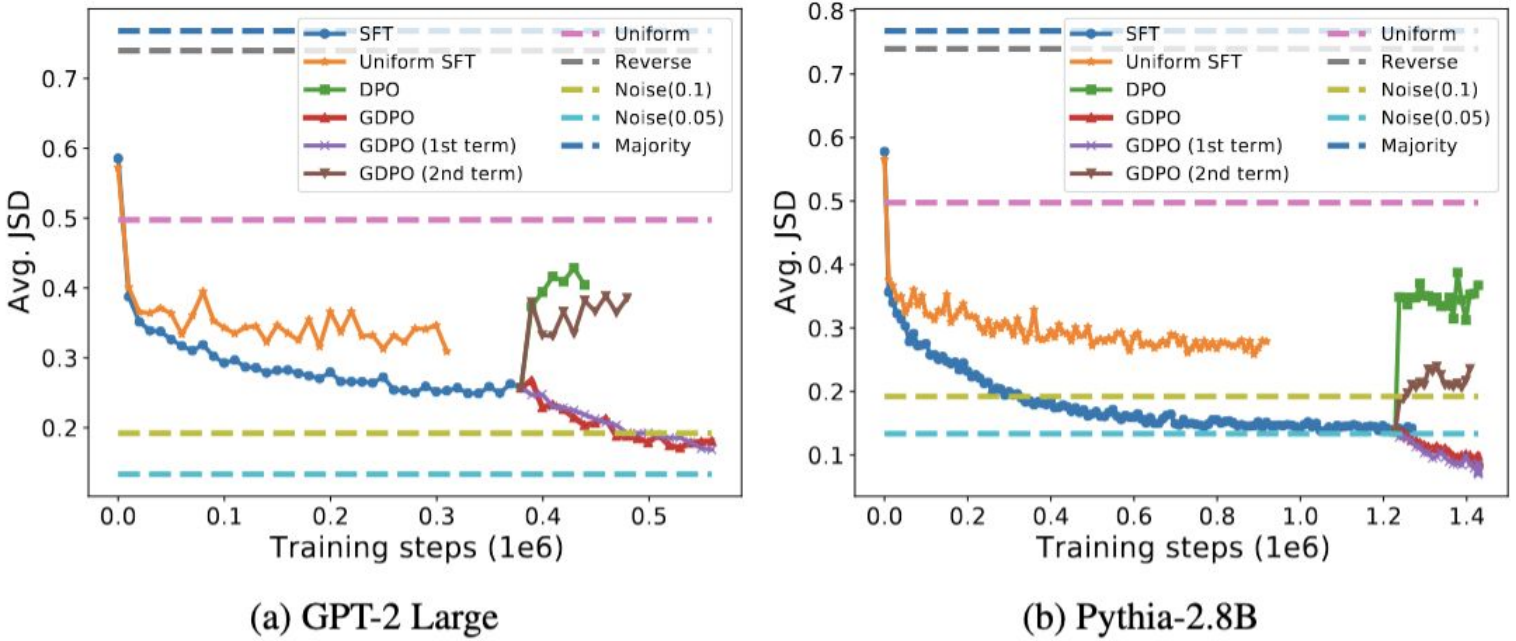


(a) GPT-2 Large



(b) Pythia-2.8B

Finding 2: GDPO Narrows the Gap from the Target Distribution



Finding 3: GDPO Excels in Controllable Opinion Generation

	Metric	GPT-2 Large					Pythia-2.8B					GPT-4o
		3-Shot	ICF	SFT	DPO	GDPO	5-Shot	ICF	SFT	DPO	GDPO	5-Shot
US	**JSD**	0.513	0.269	0.261	0.385	**0.188**	0.477	0.134	0.122	0.352	**0.068**	0.528
	CBC	0.242	0.829	0.854	0.773	**0.860**	0.248	0.973	0.987	0.899	**0.989**	0.429
	BPC	0.162	0.389	0.404	0.441	**0.627**	0.058	0.342	0.471	0.469	**0.582**	0.549
	RS	0.208	0.420	0.426	0.467	**0.479**	0.098	0.504	0.520	0.527	**0.554**	0.339
PK	**JSD**	0.530	0.307	0.263	0.370	**0.187**	0.480	0.140	0.126	0.328	**0.083**	0.552
	CBC	0.274	0.771	0.869	0.609	**0.904**	0.255	0.959	0.990	0.950	**0.991**	0.424
	BPC	0.146	0.346	0.395	0.390	**0.465**	0.066	0.403	0.435	0.387	**0.571**	0.565
	RS	0.213	0.400	0.450	0.465	**0.469**	0.111	0.542	0.539	0.540	**0.582**	0.322
SA	**JSD**	0.523	0.318	0.295	0.482	**0.185**	0.499	0.211	0.137	0.386	**0.087**	0.531
	CBC	0.248	0.700	0.836	0.742	**0.905**	0.261	0.982	0.987	0.930	**0.990**	0.445
	BPC	0.127	0.332	0.350	0.362	**0.537**	0.056	0.374	0.439	0.469	**0.588**	0.541
	RS	0.206	0.394	0.417	0.402	**0.465**	0.104	0.549	0.524	0.511	**0.536**	0.362

Table 2: Automatic Evaluation of Controllable Opinion Generation

Finding 4: GDPO Generates Minority Preferences in Testing

Methods	Model Outputs
3-Shot	<No<Semi-open-ended>Yes- for the non chance of earning one job, I pick <No>Others I would like to see employed with expertise in earth sciences. The latter 2 are a significant priority for the government...
ICF	b[2]<No>Perhaps what we need is not yet another statement of indifference, but instead, a sincere dedication to justice and a commitment to equity...
SFT	b[4]<Somewhat good job>The European Union has made some strides in tackling global climate change, but more could be done to combat climate change affirmatively.
DPO	b[3]<About right>The European Union has made great strides in reducing greenhouse gas emissions, but more must be done to tackle climate change. It's essential that they continue to take bold measures to combat this issue.
GDPO	b[1]<Very bad job>The European Union's climate policies have been a complete failure, and they need to make major reforms to make things better.
5-Shot GPT-4o	b[4]<Yes>I'm really optimistic about it! It's encouraging to see so many new job opportunities opening up and I believe it will have a positive impact on our economy
Question	*How do you evaluate the response of the European Union towards global climate change?*
Belief **Mapped Class** **Belief Distribution**	[Very good job, Somewhat good job, Somewhat bad job, Very bad job, DK/Refused] [b[5],b[4],b[2],b[1],b[0]] [0.06, 0.56, 0.24, 0.08, 0.06]

No Preference Left Behind: **Group Distributional Preference Optimization**

- Email binwei.yao@wisc.edu for any questions

Paper

Code