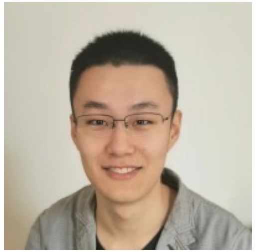
University of Virginia

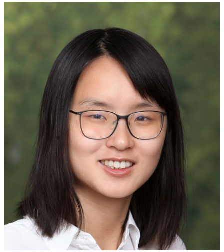
Princeton University

SimPO: Simple Preference Optimization with a Reference-Free Reward

https://arxiv.org/abs/2405.14734

Yu Meng*, Mengzhou Xia*, Danqi Chen





Neural Information Processing Systems

Aligning Language Models with Human Preferences

Preference Dataset: Signals for human desiderata

Prompt

Candidate answers

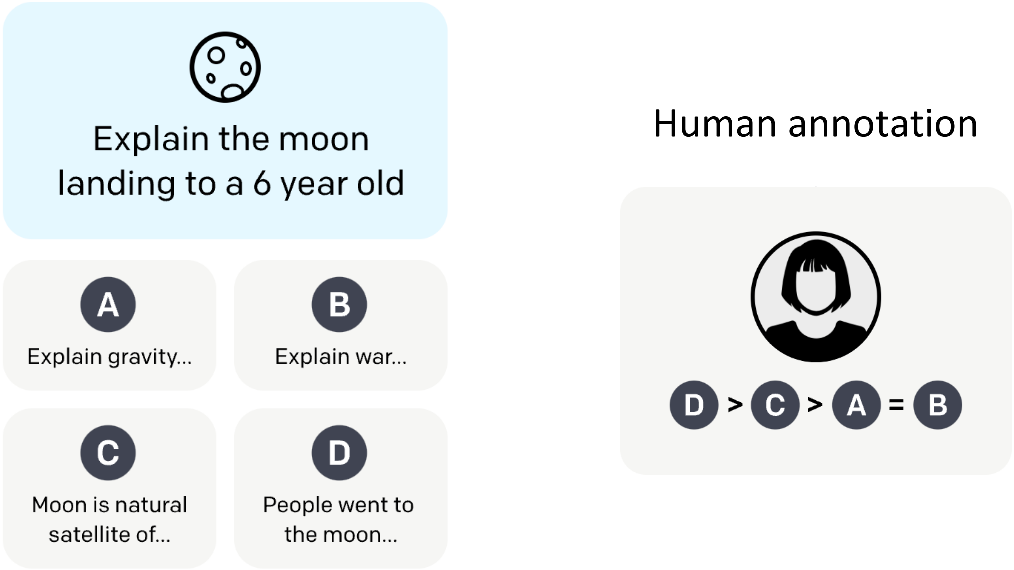
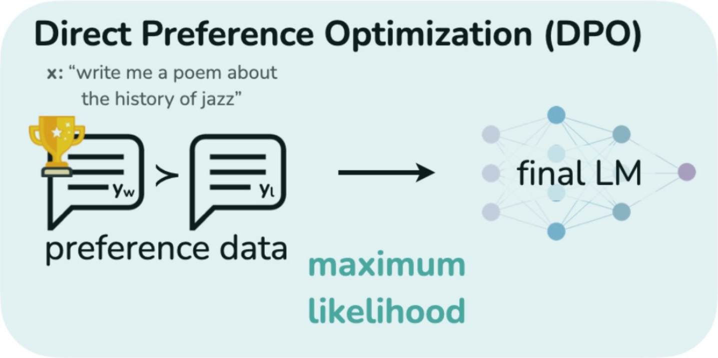


Figure from: https://openai.com/index/instruction-following/

Direct Preference Optimization (DPO)

Instead of training an explicit reward model, express reward in the form of policy model



Implicit reward expression

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

Bradley-Terry ranking objective

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

DPO objective

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

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

Discrepancy Between Reward and Generation for DPO

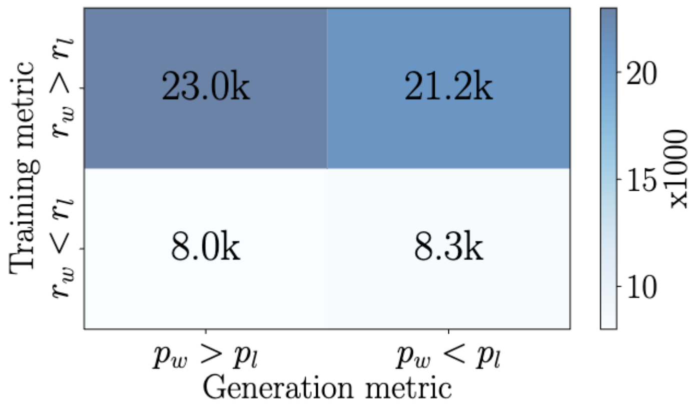
- Only policy model is used in generation

$$p_\theta(y \mid x) = \frac{1}{|y|} \log \pi_\theta(y \mid x) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_\theta(y_i \mid x, y_{<i})$$

- Reward ranking mismatches likelihood ranking

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

DPO's reward expression includes a reference model (not used in decoding)



SimPO: Length-Normalized Reward

- Consider a simple reward formulation aligned with generation

$$p_\theta(y \mid x) = \frac{1}{|y|} \log \pi_\theta(y \mid x) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_\theta(y_i \mid x, y_{<i})$$

$$\Big\downarrow \text{ scaled by constant}$$

$$r_{\text{SimPO}}(x, y) = \frac{\beta}{|y|} \log \pi_\theta(y \mid x) = \frac{\beta}{|y|} \sum_{i=1}^{|y|} \log \pi_\theta(y_i \mid x, y_{<i})$$

- No need for reference model -> better memory & compute efficiency
- Length normalization is crucial to prevent length exploitation

Introducing Target Reward Margin

- Bradley-Terry ranking objective with a margin

$$p(y_w \succ y_l \mid x) = \sigma\left(r(x, y_w) - r(x, y_l) - \boxed{\gamma}\right)$$

- Encourage a larger margin between the winning reward and losing reward

SimPO Objective

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

$$\mathcal{L}_{\mathrm{SimPO}}(\pi_\theta) = -\mathbb{E}\left[\log \sigma\left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l \mid x) - \gamma\right)\right]$$

$$r_{\mathrm{SimPO}}(x, y) = \frac{\beta}{|y|} \log \pi_\theta(y \mid x)$$

$$\Downarrow$$

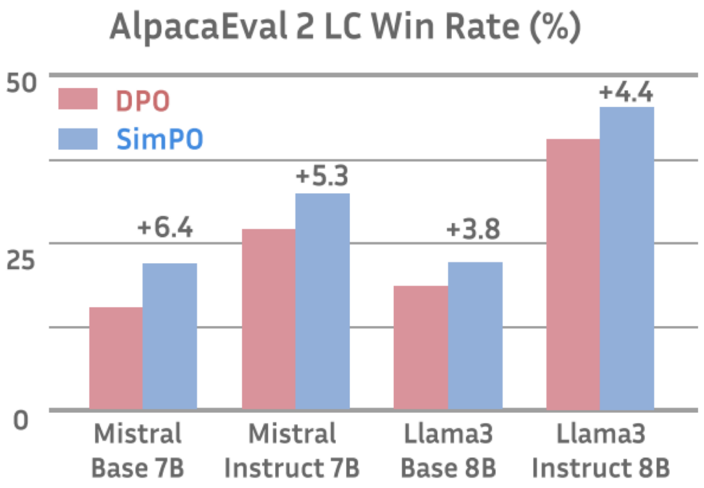
$$p(y_w \succ y_l \mid x) = \sigma\left(r(x, y_w) - r(x, y_l) - \gamma\right)$$

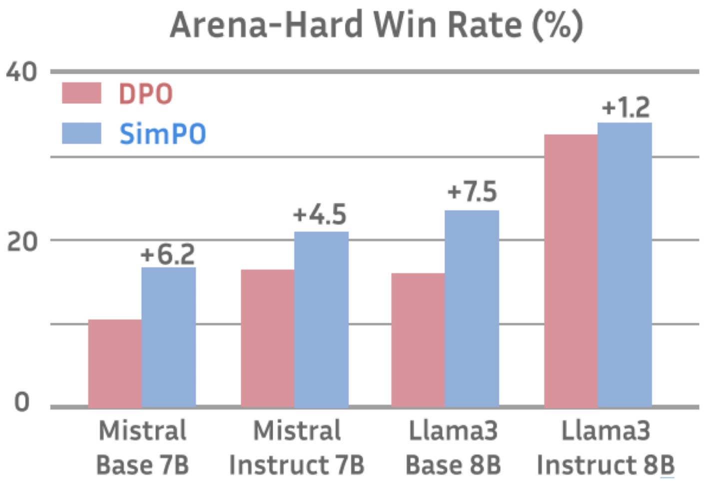
$$\Downarrow$$

$$\mathcal{L}_{\mathrm{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w|x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l|x) - \gamma\right)\right]$$

SimPO vs. DPO Results

- Mistral/Llama-3 Base = start with pretrained models, do SFT w/ UltraChat (Ding et al., 2023) + *PO w/ UltraFeedback (Cui et al., 2023)
- Mistral/Llama-3 Instruct = start with instruction-tuned models, do *PO w/ on-policy UltraFeedback data annotated w/ PairRM (Jiang et al., 2023)





Results on Gemma-2-9B

princeton-nlp / gemma-2-9b-it-SimPO like 120

AlpacaEval Leaderboard

Baseline: GPT-4 Preview (11/06)

gemma-2-9b-it:
51.1% length-controlled win rate
gemma-2-9b-it-SimPO:
72.4% length-controlled win rate

Ai2 WildBench Leaderboard V2

gemma-2-9b-it-SimPO:
1st among <10B models

Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots
(real user votes!)

Rank* (UB)	Rank (StyleCtrl)	Model
35	30	Gemma-2-27b-it
35	31	Gemma-2-9b-it-SimPO
35	33	Deepseek-Coder-v2-0724
35	33	Command R+ (08-2024)
35	35	Yi-Large
35	48	Gemini-1.5-Flash-8B-001
50	46	Command R+ (04-2024)
50	46	Qwen2-72B-Instruct
50	49	Gemma-2-9b-it

50k data
16 GPU hours
(H100)

Reward model for on-policy data annotation: ArmoRM (Wang et al., 2024)

gemma-2-9b-it-SimPO:
on-par with gemma-2-27b-it
1st among <10B models

Thank You!

Code & Models: https://github.com/princeton-nlp/SimPO

Questions? Contact us:

Yu Meng*, Mengzhou Xia*, Danqi Chen
yumeng5@virginia.edu, {mengzhou,danqic}@cs.princeton.edu