Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback

Hamish Ivison♣♠ Yizhong Wang♣♠ Jiacheng Liu♣♠
Zeqiu Wu♠ Valentina Pyatkin♣♠ Nathan Lambert♣
Noah A. Smith♣♠ Yejin Choi♣♠ Hannaneh Hajishirzi♣♠

♣Allen Institute for AI ♠University of Washington

NeurIPS 2024

HUMANE Lab 박현빈

25.03.21

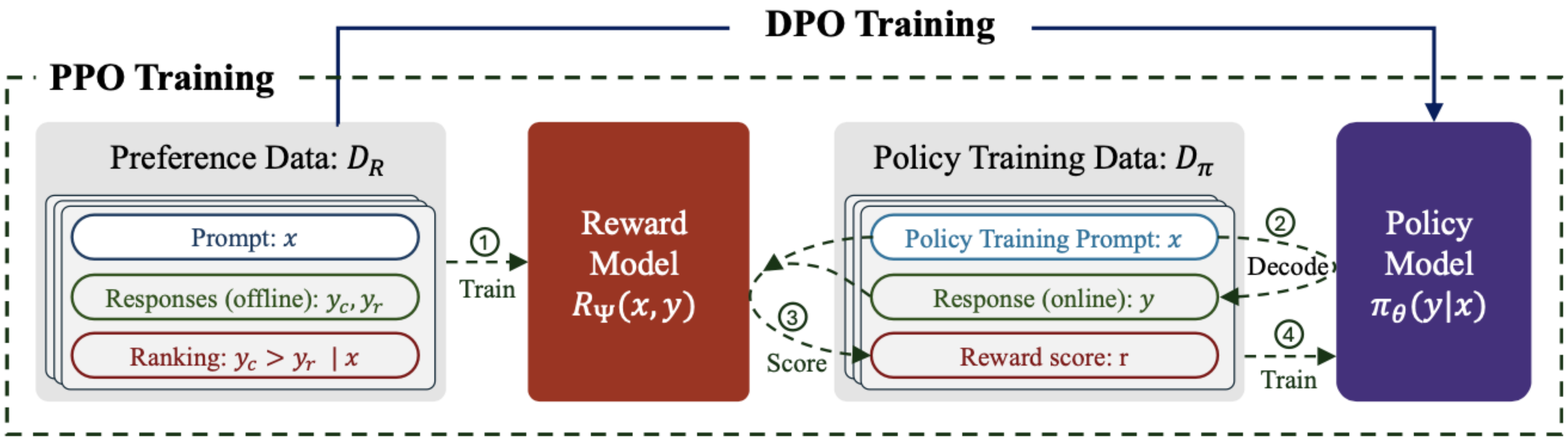
Background

- The way preference-based learning is applied varied wildly, with differing data, learning algorithms, and evaluations used, making disentangling the impact of each aspect difficult

Contribution

- Dataset
  - quality of the preferences matters more than the quality of the actual generations
- PPO vs. DPO
  - PPO outperforms DPO across varied datasets
- Reward model
  - increasing reward model size or dataset size improves reward model performance
  - however, the improvement in reward model performance does not lead to an improvement in benchmark performance
- Policy training prompts
  - using unlabelled prompts that better match the test setting during policy can further improve model performance in domain-specific settings

PPO & DPO



- PPO reward model loss function

$$\mathcal{L}_R(\psi) = -\mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}_R}\left[\log\sigma\left(R_\psi(x,y_c) - R_\psi(x,y_r)\right)\right]$$

- DPO loss function

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

- PPO policy training goal

$$\max_{\pi_\theta}\mathbb{E}_{x\sim\mathcal{D}_\pi, y\sim\pi_\theta(y|x)}\left[R_\psi(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left(\pi_\theta||\pi_{\text{ref}}\right)$$

PPO & DPO

- PPO trains on online data, DPO trains on pre-generated offline data
- DPO is more efficient in terms of compute, speed, and engineering efforts
- PPO outperform DPO

Experimental Setup

- Model
  - TULU2 13B is a series of Llama2
- Benchmark
  - factuality (MMLU)
  - reasoning (GSM8k, Big Bench Hard)
  - truthfulness (TruthfulQA)
  - coding (HumanEval+, MBPP+)
  - safety (Toxigen, XSTest)
  - instruction following (AlpacaEval, IFEval)

Preference Data

Source		# Samples	Factuality	Reasoning	Coding	Truthfulness	Safety	Inst. Following	Average
-	Llama 2 base	-	52.0	37.0	30.7	32.7	32.7	-	-
-	TÜLU 2 (SFT)	-	55.4	47.8	45.1	56.6	91.8	44.2	56.8
Web	SHP-2	500,000	55.4	47.7	40.3	62.2	90.4	45.6	56.9
	StackExchange	500,000	55.7	46.8	39.6	67.4	92.6	44.6	57.8
Human	PRM800k	6,949	55.3	49.7	**46.6**	54.7	91.9	43.4	56.9
	Chatbot Arena (2023)	20,465	55.4	50.2	45.9	58.5	67.3	50.8	54.7
	Chatbot Arena (2024)	34,269	55.7	50.4	37.7	56.7	58.1	50.7	51.5
	AlpacaF. Human Pref	9,686	55.3	47.6	43.3	56.1	90.7	44.5	56.2
	HH-RLHF	158,530	54.7	46.0	43.6	65.6	**93.1**	45.4	58.1
	HelpSteer	9,270	55.2	48.2	46.5	60.3	92.5	45.2	58.0
Synthetic	AlpacaF. GPT-4 Pref	19,465	55.3	49.1	43.4	57.7	89.5	46.3	56.9
	Capybara 7k	7,563	55.2	46.4	46.4	57.5	91.5	46.1	57.2
	Orca Pairs	12,859	55.5	46.8	46.0	57.9	90.5	46.2	57.2
	Nectar	180,099	55.3	47.8	43.2	68.2	**93.1**	47.8	59.2
	UltraF. (overall)	60,908	**55.6**	48.8	46.5	67.6	92.1	51.1	60.3
	UltraF. (fine-grained)	60,908	55.3	**50.9**	45.9	**69.3**	91.9	**52.8**	**61.0**

- The use of per-aspect annotations is more important for performance than the quality of the models used to generate completions for the dataset (HelpSteer, UltraFeedback)

DPO vs. PPO

Data / Model	Alg.	Factuality	Reasoning	Coding	Truthfulness	Safety	Inst. Foll.	Average
Llama 2 base	-	52.0	37.0	30.7	32.7	32.7	-	-
TÜLU 2 (SFT)	-	55.4	47.8	45.1	56.6	91.8	44.2	56.8
StackExchange	DPO	**55.3**	**47.8**	42.4	**56.2**	92.0	46.7	56.7
	PPO	55.1	**47.8**	**46.4**	54.2	**92.6**	**47.4**	**57.3**
ChatArena (2023)	DPO	**55.4**	**50.2**	45.9	**58.5**	67.3	**50.8**	54.7
	PPO	55.2	49.2	**46.4**	55.8	**79.4**	49.7	**55.9**
HH-RLHF	DPO	**55.2**	47.6	44.2	**60.0**	**93.4**	46.6	57.8
	PPO	54.9	**48.6**	**45.9**	58.0	92.8	**47.0**	**57.9**
Nectar	DPO	**55.6**	45.8	39.0	**68.1**	**93.3**	**48.4**	58.4
	PPO	55.2	**51.2**	**45.6**	60.1	92.6	47.4	**58.7**
UltraFeedback (FG)	DPO	55.3	50.9	45.9	69.3	**91.9**	52.8	61.0
	PPO	**56.0**	**52.0**	**47.7**	**71.5**	91.8	**54.4**	**62.2**
Avg. Δ b/w PPO & DPO		-0.1	+1.3	+2.9	-2.5	+2.3	+0.1	+0.7

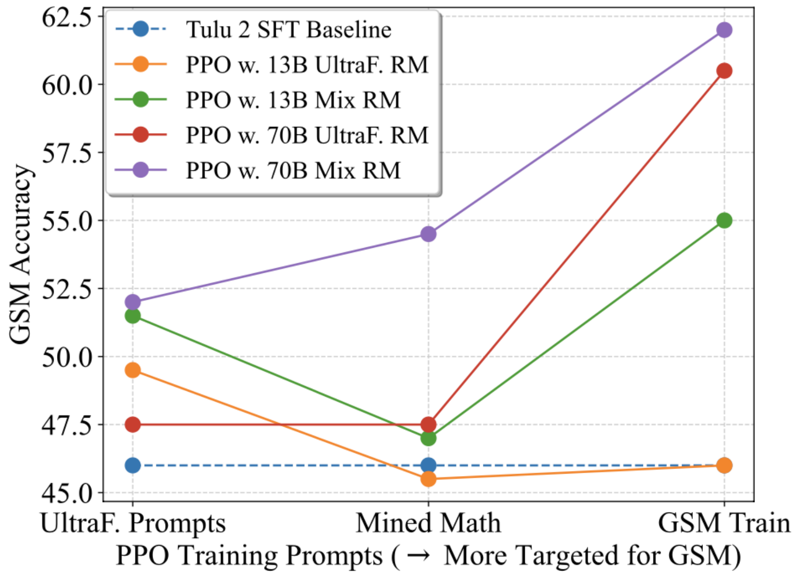
- PPO outperform DPO
- PPO-trained models are more likely than DPO-trained models to perform CoT reasoning

Reward Models

Reward Model	Direct Eval.		PPO Training Perf. (w. UltraF. prompts)		
	RewardBench Score	Best-of-N over SFT Avg. Perf. (Δ)	GSM Acc.	AlpacaEval2 winrate	Avg. on All Evals.
13B UltraF. RM	61.0	56.9 (+5.8)	53.0	26.1	62.2
13B Mix RM	**79.8**	58.3 (+7.3)	51.0	25.7	61.6
70B UltraF. RM	73.6	**61.1 (+10.3)**	**58.0**	26.7	**62.8**
70B Mix RM	73.9	60.6 (+9.5)	51.5	**31.6**	61.8

- Mix: UltraFeedback, HelpSteer, Nectar, StackExchange, HH-RLHF, PRM800k
- Either increasing the reward model dataset (‘Mix’) or reward model size (from 13B to 70B) improves direct RM performance
- Improvements in reward models result in surprisingly small improvements in overall downstream performance

Policy Training Prompts



Reward Model	Prompts	GSM %	Coding	Avg. Across All Evals
Tulu 2 SFT	-	46.0	45.1	56.8
13B UltraF.	UF	53.0	47.7	**62.2**
13B UltraF.	Mixed	**54.5**	**47.8**	61.9
13B Mix	UF	**51.0**	**46.8**	**61.6**
13B Mix	Mixed	50.5	43.8	60.9
70B UltraF.	UF	**58.0**	47.3	**62.8**
70B UltraF.	Mixed	56.5	**48.4**	62.4
70B Mix	UF	51.5	**46.1**	**61.8**
70B Mix	Mixed	**52.0**	44.9	61.1

- Larger reward models perform better when closely matching train prompts to test settings
- using mixed prompts does not seem to improve performance in the generalist setting

Conclusion

- High quality, synthetic preference dataset
- A large reward model
- Train with PPO
- collect domain-specific prompts for policy training