Data-dependent Exploration for Online Reinforcement Learning from Human Feedback

Zhen-Yu ZHANG, Yuting TANG, Jiandong ZHANG, Lanjihong MA, Masashi SUGIYAMA

Frontiers of Computer Science, DOI: 10.1007/s11704-026-60860-y

Problems & Ideas

- **Problems in online RLHF**
  - Preference feedback is collected on-the-fly, so exploration is critical for efficient alignment.
  - Existing bonuses often rely on on-policy expectations that are hard to estimate from historical data.
  - The policy may prematurely down-weight under-explored regions that contain high-value responses.
- **Idea: Data-dependent Exploration for Preference Optimization (DEPO)**
  - Assign larger exploration bonuses to under-covered directions and inject them into preference optimization.

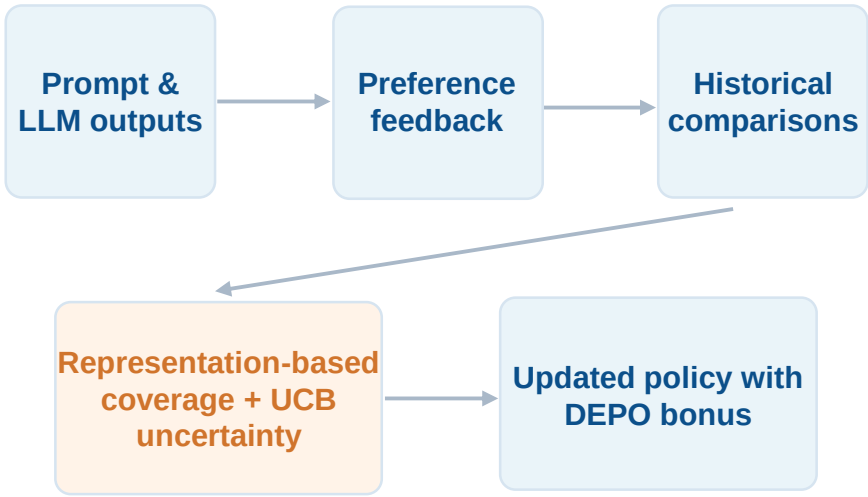


Figure 1. DEPO pipeline.

Main Contributions

- Contributions:
  - A lightweight data-dependent exploration bonus built from historical preference comparisons.
  - A representation-based uncertainty measure that gives larger bonuses to under-covered directions.
  - A data-dependent preference-regret bound that adapts to the geometry of collected data.
  - Consistent improvements over DPO, XPO, and POPO across online RLHF benchmarks and code generation.

Model	MMLU	GPQA	TruthfulQA	GSM8k	AE2	MT	IID		Alpaca	
							WR	AvgR	WR	AvgR
Llama-3-8B-SFT	62.61	32.40	53.50	71.77	10.22	7.66	–	-5.54	–	-7.92
DPO-iter1	62.87 ± 0.46	**32.84 ± 0.39**	56.24 ± 0.71	76.65 ± 0.52	–	–	62.4 ± 0.63	-4.50 ± 0.24	78.1 ± 0.58	-6.02 ± 0.16
DPO-iter2	62.92 ± 0.54	30.76 ± 0.67	57.66 ± 0.48	77.80 ± 0.36	–	–	66.6 ± 0.72	-3.59 ± 0.35	87.1 ± 0.41	-3.34 ± 0.29
DPO-iter3	63.08 ± 0.38	31.24 ± 0.74	59.61 ± 0.45	77.38 ± 0.62	36.14 ± 0.57	8.29 ± 0.34	72.5 ± 0.66	-2.36 ± 0.13	91.2 ± 0.47	-0.02 ± 0.19
XPO-iter1	62.66 ± 0.51	32.47 ± 0.43	56.19 ± 0.65	76.42 ± 0.48	–	–	62.6 ± 0.70	-4.40 ± 0.26	78.3 ± 0.61	-5.79 ± 0.15
XPO-iter2	63.13 ± 0.36	31.25 ± 0.59	58.61 ± 0.44	77.30 ± 0.73	–	–	67.3 ± 0.50	-3.28 ± 0.37	88.0 ± 0.52	-2.60 ± 0.24
XPO-iter3	63.12 ± 0.62	31.01 ± 0.55	59.34 ± 0.39	**78.41 ± 0.46**	38.23 ± 0.68	8.21 ± 0.41	72.9 ± 0.57	-2.11 ± 0.34	91.6 ± 0.35	0.64 ± 0.29
POPO-iter1	62.85 ± 0.44	32.38 ± 0.72	56.11 ± 0.50	76.88 ± 0.37	–	–	62.5 ± 0.59	-4.32 ± 0.36	80.0 ± 0.48	-5.68 ± 0.25
POPO-iter2	62.95 ± 0.53	31.72 ± 0.40	57.81 ± 0.69	77.39 ± 0.58	–	–	68.2 ± 0.45	-3.15 ± 0.22	89.1 ± 0.73	-2.45 ± 0.10
POPO-iter3	63.18 ± 0.41	31.94 ± 0.63	59.07 ± 0.56	77.63 ± 0.49	40.13 ± 0.71	8.38 ± 0.38	73.2 ± 0.54	-2.02 ± 0.16	92.4 ± 0.67	0.60 ± 0.23
DEPO-iter1	62.88 ± 0.47	32.83 ± 0.52	56.14 ± 0.64	76.55 ± 0.39	–	–	62.6 ± 0.58	-4.30 ± 0.31	80.2 ± 0.50	-5.61 ± 0.22
DEPO-iter2	63.12 ± 0.35	32.19 ± 0.68	58.26 ± 0.43	77.49 ± 0.55	–	–	68.9 ± 0.49	-3.01 ± 0.27	89.5 ± 0.44	-2.18 ± 0.19
DEPO-iter3	**63.23 ± 0.40**	32.50 ± 0.46	**59.82 ± 0.51**	78.39 ± 0.42	**41.56 ± 0.60**	**8.45 ± 0.36**	**74.1 ± 0.53**	**-1.94 ± 0.28**	**92.8 ± 0.39**	**0.66 ± 0.21**

Table 1 Main comparison across three online RLHF iterations. DEPO achieves the strongest overall performance at iter3, with consistent gains on AE/MT and improved win rates on both IID and Alpaca.