HelpSteer2-Preference: Complementing Ratings with Preferences

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, Yi Dong

zhilinw@nvidia.com

Why do we need HelpSteer2-Preference?

1. **It's unclear what the best approach for Reward Modelling is**
   a. **Bradley-Terry models:** OpenAI InstructGPT, Anthropic HH-RLHF, Meta Llama 3
   b. **Regression models:** Nemotron 4 340B Reward, RLHFlow ArmoRM 8B
2. **We need matched data for both approaches to find out**
   a. **Identical set of prompts and responses**
   b. **Collected for Purpose:** Retrofitting Regression data for preference is not sufficient
   c. **High Quality:** Garbage in; Garbage out - need to ensure high signal to noise ratio
3. **Open-source dataset to support open science**
   a. **Data Gap:** There's currently no open-source dataset that fulfills all of the criteria above

Can we create an open-source dataset to understand the best RM approach?

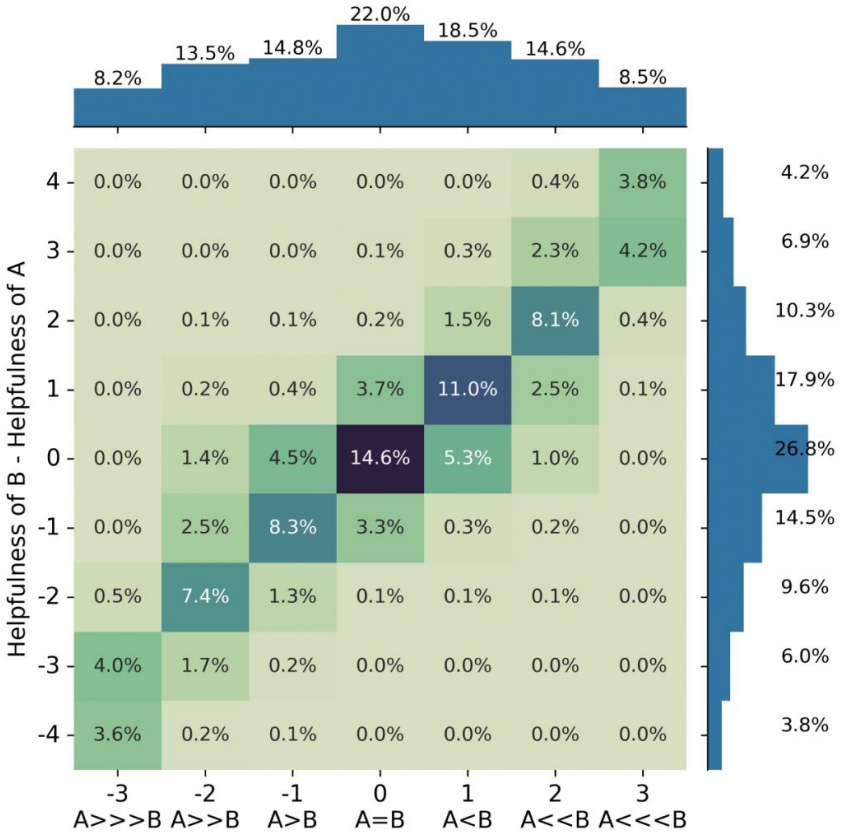
What is HelpSteer2-Preference?

HelpSteer2-Preference is an open-source, enterprise-friendly and rich dataset for **top-performing and efficient** reward modelling.

- **Top-performing:** Used to train Llama-3.1-Nemotron-70B-Reward, **No. 1 on Reward Bench (94.1)** at time of release (Oct 2024).
- **Complementary to Ratings**: Prompts and responses are identical with HelpSteer2 (which contains Likert-5 ratings of Helpfulness and other attributes), permitting fair comparison.
- **Rich data:** Each of 10k samples contains preference between two responses, preference strengths (slightly better, better and much better) and human-written preference justifications.

See https://huggingface.co/datasets/nvidia/HelpSteer2#preferences-new---1-oct-2024 with permissive CC-BY-4.0 license, where it has accumulated over 300k downloads.

HelpSteer2-Preference Data Analysis



- **Preference and Helpfulness generally correlates:** Larger helpfulness difference likely suggests stronger preference
- **Correlation not perfect**: Some samples have responses with identical helpfulness but show preference for one over the other
- **Position bias is weak:** Humans show slight preference for latter response, possibly because of recency effect. Much lower than automated evals (e.g. GPT4/Claude in MT Bench)

Reward Modelling

1. Regression and Bradley-Terry perform similarly but can complement each other to reach 94.1, which is **No. 1 on Reward Bench** (on 1 Oct 2024).
2. Optimal formulation of Bradley-Terry is **Scaled Bradley-Terry** which uses preference strength information to scale loss proportionally to strength.

Model Type	*Model*	**RewardBench** Overall	Chat	Chat-Hard	Safety	Reasoning
SteerLM Regression	HelpSteer Attributes	92.4	95.0	85.5	94.0	95.1
	Helpfulness Only	93.0	97.2	84.2	94.6	95.8
Bradley-Terry (from scratch)	Regular	91.5	97.5	80.3	90.5	97.9
	Margin	91.5	98.0	78.5	94.6	94.8
	Scaled	92.7	97.8	83.5	93.2	96.0
Bradley-Terry (init. with Helpfulness-only Regression Model)	Regular	92.7	**98.9**	82.9	93.7	95.4
	Margin	93.0	98.3	83.8	94.0	95.8
	Scaled	93.7	98.0	85.7	94.3	96.7
	Scaled + ExPO	**94.1**	97.5	85.7	**95.1**	**98.1**
External Baselines	Skywork-Reward-Gemma-2-27B	93.8	95.8	**91.4**	91.9	96.1
	TextEval-Llama3.1-70B	93.5	94.1	90.1	93.2	96.4

Using Reward Model to train Aligned Models

1. Trained Reward Model can be used with REINFORCE algorithm (RLHF) to produce **top-performing model on MT-Bench, AlpacaEval 2 LC and Arena Hard.**
2. Reward/REINFORCE models **openly accessible** (Llama 3.1 licensed) at https://huggingface.co/collections/nvidia/llama-31-nemotron-70b-670e93cd366feea16abc13d8

Model Type	*Model*	**Aligned Metrics** MT Bench (GPT-4-Turbo)	Mean Response Length (Chars.)	AlpacaEval 2.0 LC (SE)	Arena Hard (95% CI)
Offline RLHF	Regular DPO	8.66	1502.2	40.4 (1.66)	52.8 (-2.7, 2.7)
	Margin DPO	8.58	1496.6	41.1 (1.67)	52.6 (-2.7, 2.8)
	Scaled DPO	8.74	1514.8	41.0 (1.68)	52.9 (-2.4, 3.1)
Online RLHF	PPO	8.74	1842.8	43.8 (1.76)	58.6 (-2.9, 2.5)
	REINFORCE	**8.98**	2199.8	**57.6** (1.65)	**85.0** (-1.5, 1.5)
External Baselines	Llama-3.1-70B-Instruct	8.22	1728.6	38.1 (0.90)	55.7 (-2.9, 2.7)
	Llama-3.1-405B-Instruct	8.49	1664.7	39.3 (1.43)	69.3 (-2.4, 2.2)
	Claude-3-5-Sonnet-20240620	8.81	1619.9	52.4 (1.47)	79.2 (-1.9, 1.7)
	GPT-4o-2024-05-13	8.74	1752.2	57.5 (1.47)	79.3 (-2.1, 2.0)