HelpSteer2: Open-source dataset for training top-performing reward models

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, Oleksii Kuchaiev

zhilinw@nvidia.com

Why do we need HelpSteer2?

1. **Frontier models do not release alignment data**
   a. **Proprietary models:** GPT; Claude; Gemini
   b. **Open-weight models:** Llama; Mistral; Qwen
2. **Strong alignment datasets are not commercial-friendly**
   a. **GPT-4 labelled data are popular in academia:** Ultrafeedback; Nectar
   b. **Legal risks for commercial setting use**: Enterprise users are often unable to use them
3. **Commercial-friendly alignment datasets are not strong enough**
   a. Datasets (license) **:** HH-RLHF (MIT); Open Assistant (Apache 2.0); HelpSteer (CC-BY-4.0)
   b. Weaker than GPT-4-labelled datasets: Less good aligned models trained with these.

Can we create an enterprise-friendly dataset that is strong for alignment?

What is HelpSteer2?

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

- **Top-performing:** Used to train Nemotron-4-340B-Reward, **No. 1 on Reward Bench (92.0)** at time of release (Jun 2024). Also used by all Top 10 models on Reward Bench as of 8 Nov. 2024.
- **Efficient**: Contains **only 10k pairs** of model-responses to real-world prompts. Each responses is annotated for Helpfulness, Correctness, Coherence, Complexity and Verbosity on a Likert-5 Scale.

Available at https://huggingface.co/datasets/nvidia/HelpSteer2 with permissive CC-BY-4.0 license, where it has accumulated over 300k downloads in 4 months.

HelpSteer2 Collection

Using HelpSteer [1] collection strategy as a start, we implemented these key changes

1. **Stronger, more diverse models:**
   a. **HelpSteer:** 1 model (Nemotron 2 43B)
   b. **HelpSteer2:** 6 models (Nemotron 2 43B; Nemotron 3 8B/22B; Nemotron 4 15B/340B; Mixtral 8*7B)
2. **Multi-turn prompts:**
   a. **HelpSteer:** Single-turn prompts only
   b. **HelpSteer2:** Single-turn and multi-turn prompts (to support reward modeling of response to multi-turn prompts)
3. **Multiple annotator per task:**
   a. **HelpSteer:** 1 annotator per task
   b. **HelpSteer2:** 3-5 annotators per task (depending on how much initial 3 annotators disagree).

[1] https://aclanthology.org/2024.naacl-long.185/

HelpSteer2 Preprocessing

What if annotators disagree with one another?

1. **Work with vendor to make sure annotators interpret guidelines correctly**
   a. Interrater agreement (Weighted Cohen's Kappa) increases from 0.465 (moderate) to 0.706 (good) for helpfulness
2. **Remove outlier annotations at a task-level**
   a. When we have 5 annotations per task, we often observe 1 or 2 of them being drastically different from others.
   b. Taking a simple mean of all annotations introduces lots of noise (e.g. 0, 0, 0, 0, 4) → mean 0.8
   c. Instead, we identify three annotations that agree most e.g. (0, 0, 0) → mean 0
3. **Remove outlier tasks**
   a. Sometimes, even the three most-agreeing tasks disagree a lot e.g. helpfulness (1, 2, 4)
   b. To avoid noise from these tasks, we remove tasks that disagree > 2 on helpfulness.
   c. Interrater agreement increases from 0.706 (good) to 0.791 (almost excellent)

Reward Modelling

1. Nemotron 4 340B trained on HelpSteer2 scores **No. 1 on Reward Bench** (on 12 June 2024).
2. Llama 3 70B model trained on **HelpSteer2 is much better than other commercial friendly datasets**

Source of Model/Training Data	*Model*	Reward Bench Primary Dataset					Prior Sets
		Overall	Chat	Chat-Hard	Safety	Reasoning	
Proprietary Models	**Nemotron-4 340B RM** (w. HelpSteer2)*	**92.0**	95.8	**87.1**	91.5	93.7	67.4
	Cohere May 2024	89.5	96.4	71.3	**92.7**	97.7	**78.2**
	Gemini 1.5 Pro-0514	88.1	92.3	80.6	87.5	92.0	-
	Cohere March 2024	87.1	94.7	65.1	90.3	**98.2**	74.6
	GPT-4-0125-preview	85.9	95.3	74.3	87.2	86.9	70.9
	GPT-4-0409-preview	85.1	95.3	75.4	87.1	82.7	73.6
	GPT-4o-0513	84.7	**96.6**	70.4	86.7	84.9	72.6
	Claude-3-Opus-02292024	80.7	94.7	60.3	89.1	78.7	-
Trained with GPT-4 Generated Data	ArmoRM-Llama 3 8B	**90.8**	96.9	**76.8**	**92.2**	**97.3**	74.3
	RLHFlow-Llama 3 8B [33]	87.1	**98.3**	65.8	89.7	94.7	**74.6**
	Eurus RM Mistral 7B [34]	82.8	98.0	65.6	81.2	86.3	71.7
	Starling RM Yi 34B [16]	82.7	96.9	57.2	88.2	88.5	71.4
	Prometheus 2 Mistral 8*7B [36]	75.3	93.0	47.1	83.5	77.4	-
Trained with Data allowing Permissive Use	**Llama 3 70B RM** (w. HelpSteer2)*	**88.8**	91.3	**80.3**	**92.8**	**90.7**	66.5
	Llama 3 70B (w. Open Assistant)*	79.1	91.3	59.2	76.0	89.9	66.7
	Llama 3 70B Instruct [8]	76.0	**97.6**	58.9	69.2	78.5	**70.4**
	Llama 3 70B (w. HH-RLHF)*	73.9	94.4	54.6	81.2	65.6	68.8
	Pythia 1.4B (w. Open Assistant)	70.0	88.5	48.7	65.3	77.5	65.3
	Llama 3 70B (w. HelpSteer)*	66.1	93.3	59.7	56.8	54.9	67.7

Table 3: Performance of Models on Reward Bench. Higher is better for each category. All numbers except models trained by us and marked with * are taken from Reward Bench leaderboard [35].

Using Reward Model to train Aligned Models

1. **Aligned Models matches or exceeds Llama 3 70B Instruct**
2. **Only 10K human-annotated data in HelpSteer2 vs. 10M for Llama3**

Technique	*Model*	MT Bench (GPT-4-Turbo)	Mean Response Length (Chars.)	TruthfulQA MC2	AlpacaEval 2.0 LC (SE)	Arena Hard (95% CI)
Baseline	GPT-4-0613*	8.12	1057.1	0.5900	30.20 (1.07)	37.9 (-2.8, 2.4)
	Llama 3 70B Instruct*	8.16	1683.0	0.6181	**34.40** (1.38)	41.1 (-2.0, 2.2)
SFT	SFT w. DA	7.96	1514.4	0.6025	32.87 (1.40)	39.6 (-2.3, 2.4)
DPO	DPO w. HelpSteer2	8.04	1532.1	<u>0.6321</u>	30.70 (1.36)	<u>41.8</u> (-2.3, 2.3)
	Iterative DPO w. DA	8.09	1492.0	**0.6328**	29.17 (1.35)	**42.5** (-2.1, 2.4)
PPO	PPO w. HelpSteer2	8.13	1497.3	0.5629	<u>33.17</u> (1.38)	39.9 (-2.4, 2.0)
SteerLM	SteerLM w. DA	8.17	1444.1	0.5919	31.10 (1.37)	39.3 (-2.6, 2.4)
	SteerLM 2 Iter. 1 w. DA	<u>8.24</u>	1523.0	0.5911	31.10 (1.35)	38.8 (-2.3, 2.7)
	SteerLM 2 Iter. 2 w. DA	**8.28**	1471.9	0.5913	29.93 (1.35)	39.1 (-2.2, 2.4)
Ablation	SFT w. Open Assistant	6.75	676.0	0.5137	13.94 (0.82)	9.8 (-1.1, 1.4)
	SteerLM w. Open Assistant	7.44	1001.3	0.5713	20.87 (1.10)	19.2 (-2.0, 1.7)

Table 4: Evaluation of Aligned Models. Higher is better for each metric, except Mean Response Length. Because we use the Llama 3 70B Base model [8] for all aligned model experiments, we use Llama 3 70B Instruct model as a baseline, together with GPT-4-0613. Models trained "w. DA" use the Daring Anteater dataset. Metrics for models marked with * are taken from external leaderboards [52–55]. **Bold** is the top model and <u>underlined</u> is the next best.