












HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, Oleksii Kuchaiev
zhilinw@nvidia.com

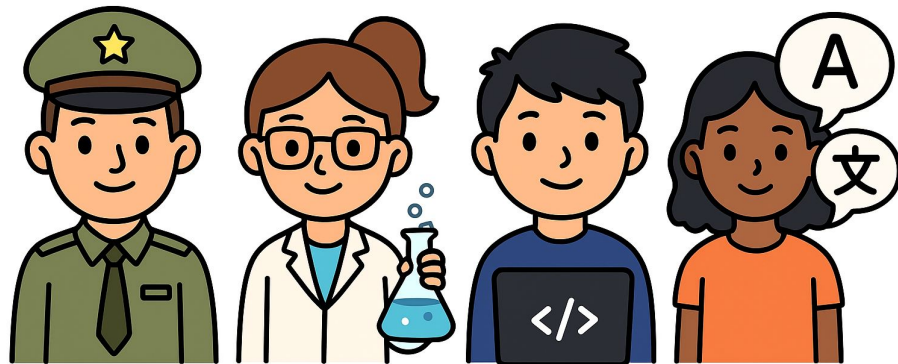
Limitations of Existing Preference Dataset

Dataset	Released	Quality	Diverse	Multilingual	Annotator	Commercial Use
HH-RLHF [1]	Apr 22	✗	✗	✗		✓
Open Assistant [2]	Apr 23	✗	✗	✓		✓
UltraFeedback [3]	Oct 23	●	●	✗		●
HelpSteer [4]	Nov 23	●	●	✗		✓
Nectar [5]	Nov 23	●	●	✗		●
Skywork-Preference [6]	Oct 24	✓	●	✗	 + 	●
HelpSteer2-Preference [7]	Oct 24	✓	●	✗		✓
INF-ORM-Preference [8]	Dec 24	✓	●	✗	 + 	●
HelpSteer3-Preference (Ours)	Mar 25	✓	✓	✓		✓

Diversity of Tasks that represent LLM Usage

Existing Preference Datasets are mainly working on General Chat settings

Other domains require expert annotators that are hard to recruit for



General

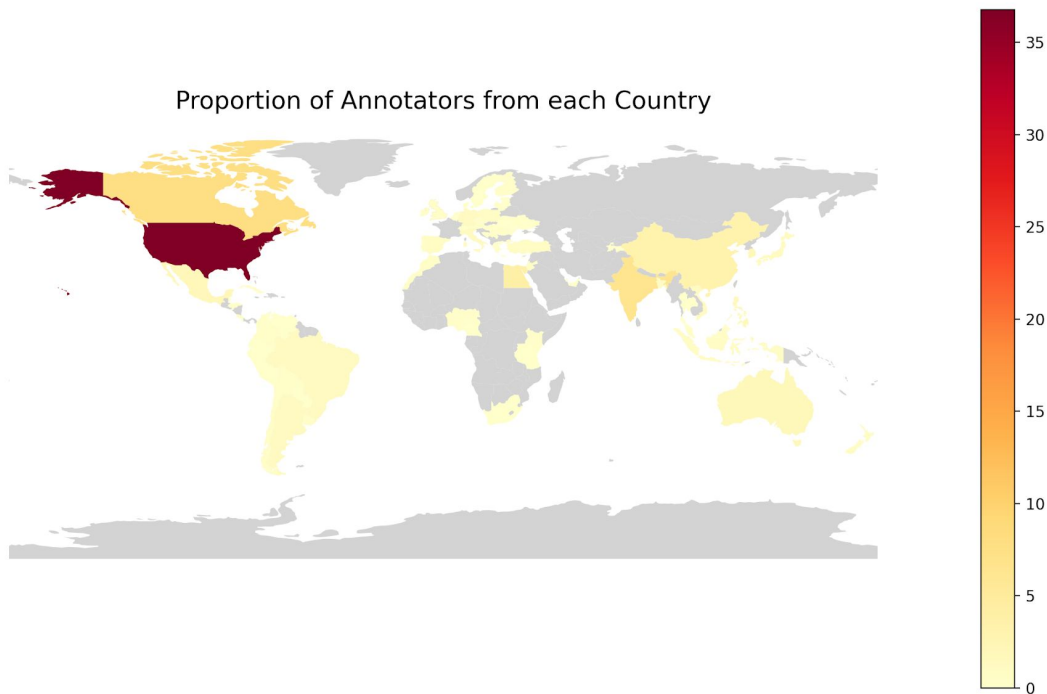
STEM

Code

Multilingual

Multi-Region Annotators to reflect Global Preferences

Annotators from more than 80 countries



Diversity of Languages across Coding and Multilingual

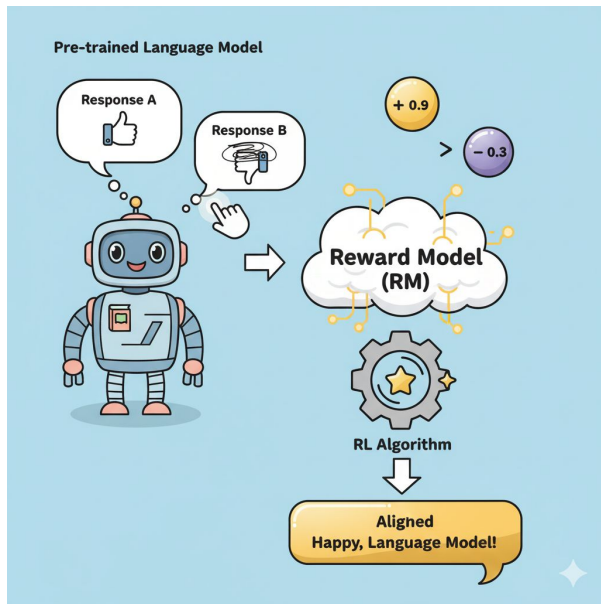
14 Coding languages + 13 Natural Languages

Empowers Preference Learning beyond English-only data in current datasets

<i>Subset</i>		% of Samples by Language												
Code	Python	JS/HTML/CSS	C#	SQL	Java	C++	Go	C	PHP	TS	PowerShell	Rust	R	Bash
Proportion (%)	38.2	23.3	5.5	5.2	5.1	5.0	3.6	3.3	3.3	3.1	1.2	1.1	1.1	1.0
Multilingual	Chinese	Korean	French	Spanish	Japanese	German	Russian	Port.	Ital.	Viet.	Dutch	Polish	Indonesian	
Proportion (%)	30.2	10.4	10.1	10.1	7.0	6.2	6.0	5.6	5.2	2.4	2.4	2.2	2.1	

Evaluation Metrics

RM-Bench for Reward Model in RLHF



JudgeBench for LLM-Judges



Comparison with SOTA Reward Models

Substantially higher (>10%) RM-Bench and JudgeBench compared to existing SOTA models.

<i>Model</i>	RM-Bench								JudgeBench				
	Chat	Math	Code	Safety	Easy	Normal	Hard	Overall	Knowl.	Reason.	Math	Coding	Overall
<i>Bradley-Terry Reward Models</i>													
English RM (General + STEM + Code)	75.4	84.5	69.3	90.4	92.1	85.7	71.1	79.9	70.8	76.5	82.1	66.7	73.7
Multilingual RM	86.2	82.4	66.8	94.1	86.5	85.4	80.0	82.4	66.2	71.4	82.1	59.5	69.4
<i>External Baselines</i>													
Llama-3.1-Nemotron-70B-Reward	70.7	64.3	57.4	90.3	92.2	76.8	48.0	70.7	62.3	72.5	76.8	57.1	66.9
Skywork-Reward-Gemma-2-27B*	71.8	59.2	56.6	94.3	89.6	75.4	50.0	70.5	59.7	66.3	83.9	50.0	64.3
Skywork-Reward-Llama-3.1-8B*	69.5	60.6	54.5	95.7	89.0	74.7	46.6	70.1	59.1	64.3	76.8	50.0	62.3

Performance of RLHF models trained with Reward Models

RLHF with Llama 3.3 70B causes it to outperform GPT-4o and Sonnet-3.5 on MT Bench, Arena Hard and WildBench

<i>Model</i>	MT Bench (GPT-4-Turbo)	Arena Hard (95% CI)	WildBench					
			Overall	Creative	Plan.	Data Analy.	Info. Seek.	Coding
Llama-3.3-70B-Instruct (Init. Policy)	8.29	62.4 (-2.5, 2.5)	52.5	55.5	54.1	48.2	54.8	51.7
+ RLOO w/ English RM	9.24	87.0 (-1.3, 1.3)	60.0	65.0	60.8	52.5	62.2	62.0
+ RLOO w/ Multilingual RM	8.81	69.8 (-1.9, 2.1)	55.5	58.7	56.9	50.8	58.4	54.7
+ RLOO w/ Baseline RM	9.04	80.7 (-1.7, 1.9)	58.9	63.6	60.9	53.4	61.9	57.6
<i>External Baselines</i>								
gpt-4o-2024-05-13	8.74	79.3 (-2.1, 2.0)	59.3	59.1	60.2	57.3	58.6	60.5
Claude-3.5-Sonnet-20240620	8.81	79.2 (-1.9, 1.7)	54.7	55.6	55.6	50.2	55.5	56.5