TOWARDS COMPREHENSIVE PREFERENCE DATA COLLECTION FOR REWARD MODELING

Yulan Hu[1,2], **Qingyang Li**[2], **Sheng Ouyang**[1], **Ge Chen**[3], **Jinman Zhao**[4], **Yong Liu**[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China [2]Kuaishou Technology
[3]University of Chinese Academy of Sciences [4]UC University of Toronto
huyulan@ruc.edu.cn, liqingyang@kuaishou.com

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) facilitates the alignment of large language models with human preferences. A critical component of RLHF is the reward model, which is trained on preference data and outputs a scalar reward for given text. However, the collection of high-quality preference data still lacks thorough investigation. Recent studies indicate that preference data is collected either by AI or humans, where chosen and rejected instances are identified between pairwise responses. We question whether this process effectively filters out noise and ensures sufficient diversity in the collected data. To address these concerns, we propose a comprehensive framework for preference data collection, decomposing the process into four incremental steps: Prompt Collection, Response Generation, Response Filtering, and Human Labeling. This framework ensures the collection of high-quality preferences while reducing reliance on human labor. We conducted comprehensive experiments using the data collected at different stages, demonstrating the effectiveness of the proposed framework.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) Ziegler et al. (2019) has demonstrated significant potential in aligning Large Language Models (LLMs) with human preferences Ouyang et al. (2022); Touvron et al. (2023); Rafailov et al. (2024). In RLHF, a reward model (RM) is typically used to output rewards for a given prompt and response, further guiding the reinforcement learning process. The RM generally relies on collected preference data for training, enabling it to distinguish between chosen and rejected responses Wang et al. (2024a).

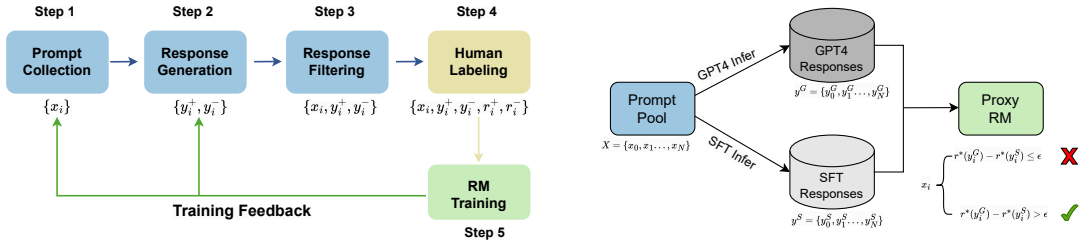
Recent years have seen increasing discussions on constructing and improving RMs. Several attempts involving the design of model architectures or learning criteria have been developed Artetxe et al. (2022); Zhang et al. (2024); Wang et al. (2024b), while studies on how to construct high-quality data have been relatively overlooked. As revealed in Wang et al. (2024a), the preference data used for RM training—whether off-the-shelf or collected by AI or humans—often contains noise and may not be suitable for RM training. To obtain preference data suitable for RM training, it is important to investigate how to properly construct the preference data procedure.

In this paper, we present the first comprehensive study on collecting high-quality preference data for RM training. Specifically, we decompose the preference data collection process into four sub-steps: **Prompt Collection**, which collects challenging prompts that the base model struggles to handle; **Response Generation**, which produces diverse responses to enhance the model's generalization; **Response Filtering**, which removes noisy response pairs; and **Human Labeling**, which further review the preference among the pairs. Finally, the RM is trained on the data reviewed by human labelers. The proposed framework represents a preliminary attempt to thoroughly investigate preference data collection for RMs. Compared to relying solely on AI or human annotation, this framework effectively aligns the collected data with human preferences while significantly reducing the amount of human labor required. We conducted experiments on preference data collected at different stages, demonstrating that performance improves as the quality of the preference data increases.

2 METHODOLOGY

We present the proposed framework in this section. As illustrated in Figure 1(a), the framework comprises five hierarchical steps, with the first four dedicated to preference data collection. Our main contributions focus on the first three steps, while the fourth is primarily carried out by human labelers. In the following sections, we introduce the proposed framework step by step.

*Corresponding author, liuyonggsai@ruc.edu.cn.



(a) The overview of the proposed framework.

(b) Step 1: the collection of challenging prompts.

Figure 1: The overview of the proposed framework and the collection strategy of prompts.

2.1 The Proposed Framework

We present the details of the first three steps in this subsection.

Step 1: Prompt Collection. This step aims to collect a sufficient number of challenging prompts, which will be used to generate responses for RM training. The model trained after the RLHF stage should preserve the overall capabilities of the SFT model while also being capable of handling prompts that are difficult for the SFT model. To achieve this, two critical principles should be upheld. First, the prompts should exhibit diversity to avoid the barrel effect, i.e., where the RM scores accurately in one domain while being less precise in another. Second, the prompts should be difficult for the SFT model to handle.

Following the two principles, we develop a comprehensive strategy illustrated in Figure 1(b). We randomly sample a sufficient number of prompts from different categories to form the prompt pool $X = \{x_0, x_1, \ldots, x_N\}$, thereby fulfilling the first requirement. For the latter requirement, our core premise is that if the quality of the response inferred by the SFT model is close to that of strong LLM models, e.g., GPT-4, for the same prompt, it indicates that the SFT model can effectively resolve this prompt, eliminating the need for further fine-tuning in RLHF stage. We achieve this with the assistance of an off-the-shelf proxy RM, $r^\star$. Specifically, we first let the SFT model and the GPT-4 model generate two responses for each prompt in X, resulting in y^G and y^S. Then, we use $r^\star$ to score the (prompt, response) pairs as follows:

$$\Delta_{(y^G, y^S)} = \begin{cases} r^\star(x, y^G) - r^\star(x, y^S) \leq \epsilon, & \text{drop } x \\ r^\star(x, y^G) - r^\star(x, y^S) > \epsilon, & \text{keep } x \end{cases} \tag{1}$$

ϵ denotes the preset threshold difference between y^G and y^S. Equation 1 indicates that we only keep the prompts if the corresponding response generated by the SFT model significantly lags behind those of well-performing models. We denote the refined prompt set as X^*.

Step 2: Response Generation. RM training accepts a prompt x and two preference responses (y^+, y^-) as input. Analogous to Step 1, the quality and diversity of the responses need to be ensured. We tackle the two requirements by employing several strong LLMs, e.g., the open-source Qwen 2.5 Yang et al. (2024) and the Llama 3.1 Dubey et al. (2024) series models, as well as the GPT-4 API, to generate responses for each prompt x. Assuming we obtain k responses using different LLMs, these responses form the candidates pool, $Y_x = \{y_1, y_2, \ldots, y_k\}$. Then, we use the proxy RM r^* to score each response and obtain the score for each response. After that, we select two samples from Y_x with a certain degree of difference as a preference pair. In this way, we have preliminarily completed the construction of the preference training data, obtaining the training candidate set $\mathcal{D}$.

Step 3: Response Filtering. The training instance collected in Step 2 is formed as a triad (x, y^+, y^-). Ideally, y^+ should exhibit a certain degree of superiority over y^-, meaning that (x, y^+, y^-) should not be too easy or too hard for learning. However, such a condition cannot always be fulfilled. The proxy RM r^* is fine-tuned on fixed preference data, which inevitably exhibits a data distribution shift compared to the constructed preference dataset $\mathcal{D}$, causing bias in scoring Skalse et al. (2022).

Therefore, we further refine $\mathcal{D}$ by prompting GPT-4 to help filter out disqualified samples. Specifically, we first design different scoring criteria for prompts belonging to different categories. Within each criterion, each sample is evaluated on five levels, where 1 represents the worst and 5 represents the best. Then, we employ GPT-4 to

score each sample in $\mathcal{D}$, obtaining the score pairs as (x, y^+, r^+) and (x, y^-, r^-). Based on these scores, we consider that two kinds of samples should be discarded, as shown in Table 1. First, responses with identical scores, as such pairs cannot provide any discriminative knowledge during RM training. Second, responses that exhibit extreme distinctness, for example, pairs where r^+ is scored 5 and r^- is scored 1. We consider that the RM possesses the ability to distinguish samples with significant divergence, thus eliminating the need for further assessment by the annotators. We denote the preference data after Step 3 as $\tilde{\mathcal{D}}$.

$< r^+, r^- >$	1	2	3	4	5
1	1-1	1-2	1-3	1-4	1-5
2	2-1	2-2	2-3	2-4	2-5
3	3-1	3-2	3-3	3-4	3-5
4	4-1	4-2	4-3	4-4	4-5
5	5-1	5-2	5-3	5-4	5-5

Table 1: The filtering matrix. The scoring pairs highlighted in green are preserved, while those in grey are discarded.

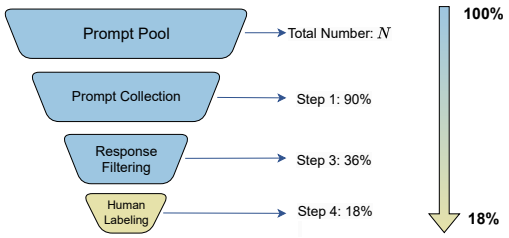


Figure 2: The data funnel illustrates the loss rate at each step. The ultimate data retention is roughly 18%.

2.2 Data Funnel

In Step 4, we let the annotators further review $\tilde{\mathcal{D}}$ obtained from Step 3. For each training sample (x, y^+, y^-), the annotators evaluate the preference between y^+ and y^- given x. The samples that do not align with the preference order in Step 3 will be discarded. We denote the ultimate dataset after Step 4 as $\mathcal{D}^*$, which will be used for RM training. The proposed framework progressively filters out the disqualified samples in a hierarchical manner. Figure 2 illustrates the data filtering funnel according to the practical collection process. Assuming that the number of prompt pools is N, Step 1 filters out 10% of easy prompts, with 90% of the prompts reserved. In Step 3, the filtering rate is comparatively large, with only 36% reserved. Finally, after human labeling, around 18% of samples are kept for RM training.

3 Experiment

3.1 Setups

We employ two SFT models built upon Llama2-13B and Llama2-70B for RM training. The preliminary prompts are collected from two sources: available open-source preference data (mainly collected from Dong et al. (2024)) and self-constructed prompts by AI or humans, each accounting for about 300,000 samples. We sample roughly 150,000 prompts from the above 600,000 samples as the initial prompt pool, then proceed with the proposed preference data collection procedure. Specifically, we train two RMs based on the data collected in Step 3 and Step 4, respectively. Approximately 54,000 preference data are collected in Step 3, and around 30,000 preference data are used in Step 4.

We save the checkpoint of the last iteration for evaluation. The RMs are evaluated from two aspects: accuracy on preference benchmarks and the BoN results. The preference benchmarks are twofold. On one hand, we evaluate the RMs on benchmarks used in Touvron et al. (2023); Bai et al. (2023), including Anthropic Helpfulness Bai et al. (2022), OpenAI Summarize Stiennon et al. (2020), OpenAI WebGPT Nakano et al. (2021), and Stanford SHP Ethayarajh et al. (2022). On the other hand, we evaluate the RMs on RewardBench Lambert et al. (2024), which contains 4 categories.

3.2 Results

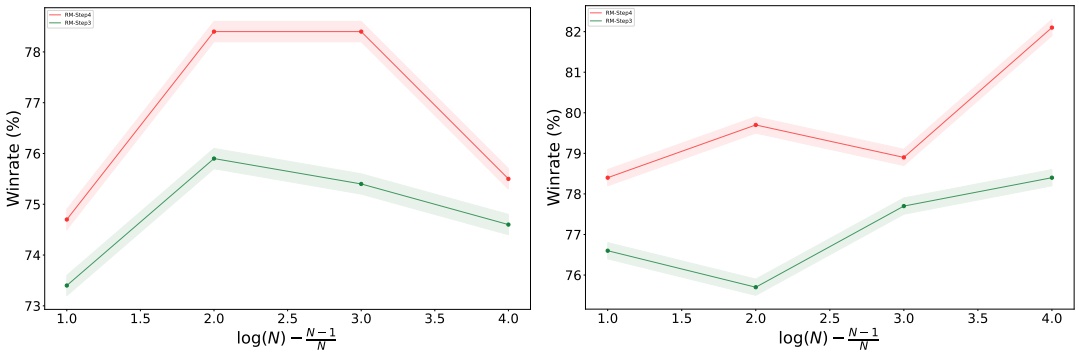
Results on preference benchmarks. We report the results on preference benchmarks in Table 2. The results for both the 13B-size and 70B-size RMs validate the improvement from step 3 to step 4, indicating that refinement of preference data can indeed boost performance. Despite the scale of preference data used in step 3 being almost twice that used in step 4, we observe that the refinement of data quality is beneficial.

Results of BoN experiments. In addition, we evaluate the RMs with the BoN policy. BoN is an inference-time sampling strategy that aims to select the answer with the highest reward from n candidates. The prompts are from AlignBench Liu et al. (2023). The gains obtained by BoN are approximated by $\log(N) - \frac{N-1}{N}$ Beirami et al. (2024). For each prompt in AlignBench, we use Llama2-13B SFT model to generate n answers and choose

Table 2: The results on preference benchmarks.

RMs		Open Preference Datasets				RewardBench				Avg
		Anthropic Helpful	OpenAI Summ.	Stanford SHP	WebGPT	Chat	Chat Hard	Reasoning	Safety	
13B	Step3	68.7	68.2	67.2	65.9	89.7	61.2	75.8	88.2	73.1
	Step4	69.6	68.6	68.1	66.7	89.1	68.2	91.8	79.9	75.3
70B	Step3	69.9	68.7	71.5	71.4	86.9	63.2	87.5	79.1	74.8
	Step4	71.4	71.4	72.1	70.8	88.8	65.6	87.7	76.8	75.6

the best answer from the answer set based on the RM score. The value of n is chosen from {5, 10, 20, 50}. We then calculate the win rate for the RM trained on preference data collected in Step 3 against Step 4, and plot the results in Figure 3. The results validate that the RMs consistently help select better answers than the SFT model for both the 13B and 70B models, further verifying the enhancement in performance with the refinement of data.



(a) 13B (b) 70B

Figure 3: The win rates of the RM trained with preference data in Step 4 against Step 3.

4 Conclusion

In this work, we propose a systematic framework for preference data collection tailored to RM training. By decomposing the collection pipeline into distinct sub-steps, our approach facilitates the acquisition of high-quality preference data while minimizing human annotation effort. We empirically validate the framework through evaluations on preference data benchmarks and downstream policy learning tasks, showing notable improvements in data quality. As an initial exploration, we believe this framework addresses critical gaps in existing practices and contributes a solution to the broader LLM community.

References

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. Efficient large scale language modeling with mixtures of experts, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pp. 5988–6008. PMLR, 2022.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024a.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble. *arXiv preprint arXiv:2401.16635*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.