Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ETH AI CENTER

ETH AI Center Project Spring 25

Active Learning for Sample-Efficient RLHF

Marian Schneider[1], Jessica Lam Jia Hong[2], Davit Melikidze[1], Martin Wertich[1], Barna Paztor[3], Ido Hakimi[3]
[1]ETH Zurich, D-INFK [2]Institute of Neuroinformatics, UZH and ETH [3]ETH AI Center, ETH Zurich

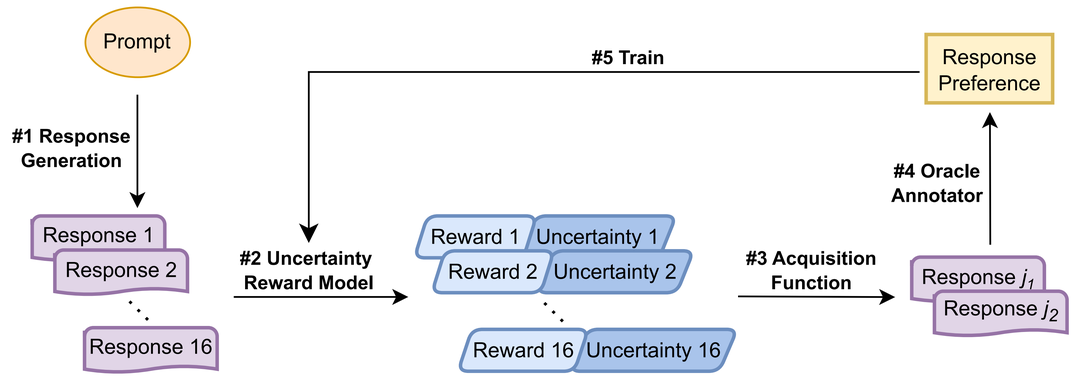

Figure 1: Overview of our active learning-infused pipeline for preference data generation.

1. Abstract

Reinforcement Learning from Human Feedback (RLHF) is essential for aligning Large Language Models (LLMs) with human preferences and values, but **creating the preference datasets required for training is challenging** because it involves recruiting people to annotate data manually. The UltraFeedback dataset generation pipeline (Cui et al., 2024) successfully reduced this reliance on human labor by using LLM annotators to annotate various LLM responses instead. Still, the large number of LLM calls it requires can be very costly. This motivates our project - using active learning to enhance UltraFeedback's annotation efficiency. In contrast to UltraFeedback, which annotates four responses per prompt, we train an **Uncertainty Reward Model** (Osband et al., 2023) to predict the utility of the responses, then use **Double Thompson Sampling** (Wu and Liu, 2016) to acquire the most useful pair of responses per prompt and annotate them only. Our results show the effectiveness of our pipeline: **Models trained on our data achieve 0.631 on RewardBench (Lambert et al., 2024) and 0.705 on IFEval (Zhou et al., 2023)**, which is comparable to models trained on UltraFeedback data while requiring only half as many annotations.

2. Related work

Cui et al. (2024) and Lambert et al. (2025) present similar pipelines for generating RLHF preference datasets, but both sample the responses to annotate randomly, whereas we explore using active learning for more efficient acquisition. More specifically, we look to Double Thompson Sampling (Wu and Liu, 2016) as employed by Dwaracherla et al. (2024) for LLM alignment, with reward and uncertainty represented by Epistemic Neural Networks (ENNs, Osband et al. (2023)).

3. Methods

The core of our pipeline is a loop that iteratively takes prompts as input and outputs preference data, see Figure 1 for an overview.

#1 Response Generation For each prompt in the batch, we call multiple LLMs to each generate a response to the prompt.

#2 Uncertainty Reward Prediction An ENN predicts the reward (and associated uncertainty) of the responses for each prompt.

#3 Response Acquisition We identify the two responses per prompt that should undergo preference annotation via Double Thompson Sampling (Wu and Liu, 2016). For each response, a reward is sampled uniformly between the lower and upper bounds predicted by the reward model. The response with the highest reward is selected, and a second (different) response is selected by repeating this process a fixed number of times and using random sampling as a fallback.

#4 Preference Annotation An LLM serves as an oracle in determining which of the two acquired responses to choose and to reject.[1]

#5 Uncertainty Reward Model Training This batch of preference data is added to a replay buffer for training the ENN.

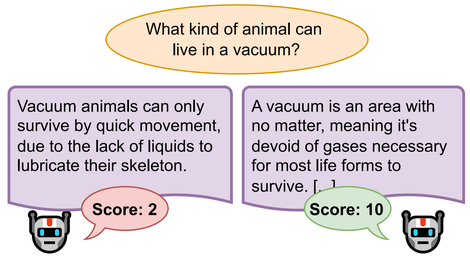

Figure 2: An example preference sample (prompt in orange, responses in purple).

We run our pipeline on 60K prompts[2] from the UltraFeedback dataset (Cui et al., 2024), so our resulting preference dataset contains 60K samples. See Figure 2 for an example. For response diversity, we use sixteen LLMs of sizes between 135 million and 72 billion from different model families (namely SmolLM, Qwen 2.5, Phi 4, Gemma 3, and Llama 3) as response generators.

4. Results

We evaluated the quality of our dataset generated by our active learning pipeline for RLHF in two ways: (i) training a reward model on it and evaluating on RewardBench, and (ii) training allenai/Llama-3.1-Tulu-3-8B-SFT using Direct Preference Optimization Rafailov et al. (2024) and evaluating on IFEval (Zhou et al., 2023).

	RewardBench	IFEval
Base (SFT) model	-	0.689
AllenAI UltraFeedback	0.747	0.735
Random	0.573	0.695
Our UltraFeedback	0.617	0.690
Active Learning	**0.631**	**0.705**

Table 1: Comparison of model performance on RewardBench and IFEval.

Based on these scores, we conclude that active learning benefits the pipeline, given that it is able to outperform the traditional UltraFeedback both in terms of score and sample efficiency.
Due to time and resource limitations, we had to simplify the oracle from UltraFeedback, only giving one overall score instead of multiple fine-grained scores. We found that our simplified version matches the original UltraFeedback oracle 44% of the time, disagrees in 20% of cases, and assigns equal scores in the remaining 36%. We attribute the discrepancy in the scores between our UltraFeedback dataset and the AllenAI UltraFeedback dataset to this change in oracle.

5. Conclusions

We can conclude that our Active UltraFeedback pipeline outperforms both random sampling and UltraFeedback on RewardBench and shows better IFEval performance with higher sample efficiency.
In future, we want to predict the reward for the pair (prompt, surrogate model) instead of (prompt, completion) to accelerate the pipeline by removing the costly response generation step.

References

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for llms, 2024. URL https://arxiv.org/abs/2402.00396.

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

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. URL https://arxiv.org/abs/2403.13787.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks, 2023. URL https://arxiv.org/abs/2107.08924.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits, 2016. URL https://arxiv.org/abs/1604.07101.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[1]Due to time and resource limitations, we simplify the preference annotation procedure used in UltraFeedback (Cui et al., 2024) and also call LLaMa 3.3 70B Instruct (Grattafiori et al., 2024) instead of GPT-4 (OpenAI, 2024).
[2]To avoid the licensing issues with the original UltraFeedback Cui et al. (2024), we use the version released by AllenAI at huggingface.co/datasets/allenai/ultrafeedback_binarized_cleaned. Both are MIT licensed.