Delft University of Technology

CSE3000 Research Project

The Human Factor: Addressing Diversity in Reinforcement Learning from Human Feedback

Javier Páez (j.paezfranco@student.tudelft.nl)

Supervised by Antonio Mone and Luciano Cavalcante

1. Introduction

Reinforcement Learning from Human Feedback (RLHF) enables training a Reinforcement Learning (RL) agent with human feedback, instead of explicitly defining a reward function (Figure 1). RLHF has been very successful, particularly in fine-tuning large language models (LLMs). However, it faces significant challenges, mainly by treating diversity as noise, overlooking human diversity, and introducing social biases [2, 3]. To mitigate this issue, recent approaches like Safe RLHF [1] have been developed.

Despite these advancements, there is no comprehensive evaluation of the true effect and relevance of conflicting data. We lack a thorough understanding of how diversity influences the overall objective.

Research question: *How can RLHF deal with possibly conflicting feedback from multiple individuals?*

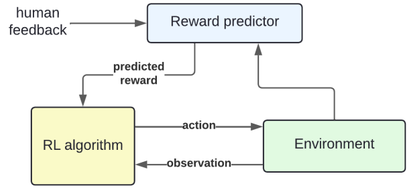

Figure 1: Schematic Illustration of RLHF

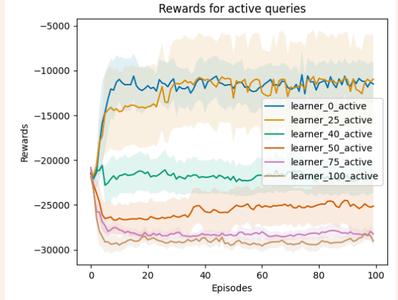

Figure 2: Results on the Pendulum environment

2. Methodology

The trajectories are generated using the Proximal Policy Optimization (PPO) algorithm. Next, trajectories are selected for human review, either randomly or through active selection (prioritizing those with the highest variance in rewards from the learned model). Feedback is then obtained based on the following equation:

$$\mu = \sigma(\sum \hat{r}(o_t^1, a_t^2) - \sum \hat{r}(o_t^2, a_t^2))$$

Where μ is the preference distribution of trajectory 1 over trajectory 2, σ is the logistic function, and r is the reward model. It is calculated synthetically by applying the softmax distribution to the difference in the sum of rewards for each trajectory.

We introduce a conflicting probability p:

$$\mu_c = \begin{cases} \mu & \text{if } U \geq p \\ 1 - \mu & \text{otherwise} \end{cases}$$

Where $p=0$, the model reduces to the original. At $p=1$, all the preferences are reversed.

We tested six different levels of conflicting data: 0%, 25%, 40%, 50%, 75%, and 100%. 0% serves as a baseline with no conflict. 25%, 40%, and 50% introduce moderate conflict, where feedback inconsistently aligns with the true reward. 75% and 100% represent challenging scenarios, greatly deviating from the true reward.

We conducted the experiments in 3 different environments of increasing complexity: Pendulum, Lunar Lander, and Bipedal Walker (Figure 3).

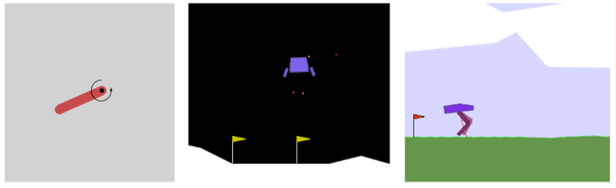

Figure 3: The environments used in the experiments. From left to right: Pendulum, Lunar Lander, and the Bipedal Walker environments.

3. Experimental Setup

We start by learning the RLHF's reward models using an ensemble of 3 predictors, followed by training the PPO agent. After 1000 environment steps, the agent policies are evaluated across 10 test episodes in an independent evaluation environment. The results are averaged across 3 experiments with different seeds.

We compare the different levels of conflicting feedback based on the mean evaluating reward, and conduct permutation tests (significance level of p=0.005) to directly compare agent performance.

4. Experiments

The results show that, in simple environments like Pendulum, RLHF can handle low levels of diversity, such as 25% (Figure 2). However, in more complex environments like Lunar Lander and Bipedal Walker, even a small of conflicting feedback rapidly degrades performance (Figures 4 & 5). Additionally, random selection seems to manage diversity better than active selection in complex tasks (Figure 4), although this improvement is insufficient to prevent performance decline.

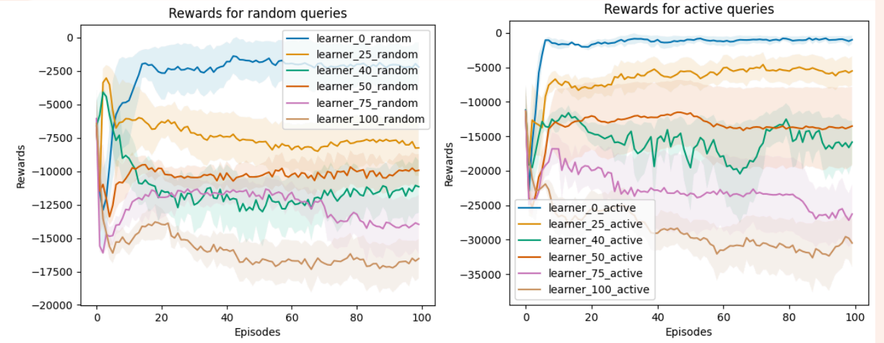

Figure 4: Results on the Lunar Lander environment

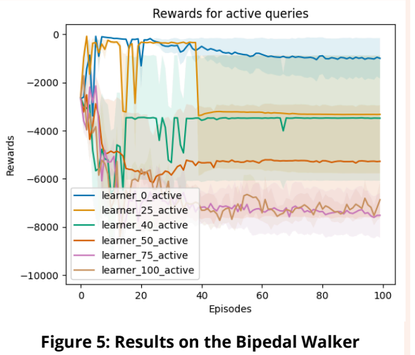

Figure 5: Results on the Bipedal Walker environment

5. Conclusion

Our results show that the performance of RLHF is significantly impacted by even modest amounts of conflicting feedback, with degradation observed at levels as low as 25%. Only in extremely simple environments like Pendulum can RLHF barely maintain its performance. Moreover, we found that randomly selecting queries yields better results than active learning under high feedback diversity.

We hope that our work stimulates investigation into alternative reward models and query selection strategies.

References

[1] J. Dai, et al., "Safe rlhf: Safe reinforcement learning from human feedback"

[2] S. Casper, et al., "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback"

[3] S. Chakrabort, et al., "MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences"