The Role of Feedback Variety in Reinforcement Learning from Human Feedback

Ivan Makarov, Luciano Cavalcante Siebert, Antonio Mone

EEMCS, Delft University of Technology

Introduction

Reinforcement Learning from Human Feedback (RLHF) [1] is a variant of reinforcement learning (RL) that learns from human feedback instead of relying on a predefined reward function. RLHF typically involves two main phases: reward learning and RL training. During reward learning, human feedback is used to train a reward model, which is then employed in conventional RL algorithms. Various feedback types can be used, including numeric scores, rankings, and corrections. This research provides an empirical comparative analysis of different feedback types in RLHF systems, highlighting trade-offs and demonstrating how various feedback types found in the literature can be practically implemented.

Background

To train the reward model with a given feedback type, an appropriate loss function must be specified. We opted for trajectories (multiple state-action pairs) as our feedback granularity and selected three representative feedback types for this research.

Scalar Feedback [2] involves a human teacher assigning numerical ratings to trajectories. The precision of Scalar Feedback comes at a cost, as it's challenging for humans to quantify rewards accurately and demands more cognitive effort [3]. The loss function is the Mean Squared Error (MSE) between the predicted reward and the scalar value:

$$L^{\mathrm{MSE}}(\theta, D) = \frac{1}{|D|} \sum_{(\sigma^i, y) \in D} \left(y - \hat{r}_\theta(\sigma^i)\right)^2 \tag{1}$$

Preference Feedback [1] requires indicating the preferred trajectory from a pair. This is the most popular feedback type for RLHF [3]. A preference predictor using the reward model $\hat{r}_\theta$ is defined as:

$$P_\theta(\sigma^1 > \sigma^0) = \mathrm{sigmoid}\left(\sum_t \hat{r}_\theta(s_t^1, a_t^1) - \sum_t \hat{r}_\theta(s_t^0, a_t^0)\right) \tag{2}$$

We update $\hat{r}_\theta$ by minimizing the standard binary cross-entropy objective:

$$L^{\mathrm{BCE}}(\theta, D) = -\frac{1}{|D|} \sum_{(\sigma^0, \sigma^1, y) \in D} \left((1-y) \log P_\theta(\sigma^0 > \sigma^1) + y \log P_\theta(\sigma^1 > \sigma^0)\right) \tag{3}$$

Preference Feedback is easier for humans to provide than Scalar Feedback, but it conveys less information and only indicates trajectory preference [3].

Marginal Preference Feedback [4] conveys more information than Preference Feedback by quantifying the extent of preference for one trajectory over another. The idea is to include a margin to use additional information on preference intensity, training the reward model to assign distinct scores to trajectories with greater differences. The loss function remains unchanged, but the preference predictor is updated to include the margin:

$$P_\theta(\sigma^1 > \sigma^0) = \mathrm{sigmoid}\left(\sum_t \hat{r}_\theta(s_t^1, a_t^1) - \sum_t \hat{r}_\theta(s_t^0, a_t^0) - m(r)\right) \tag{4}$$

The margin m(r) is a discrete function of the preference rating, with larger margins for higher preference intensity and smaller margins for lower intensity.

Methodology

The Imitation library [5] provides a framework for training reward models with RLHF. It implements only Preference Feedback, using synthetic feedback instead of human feedback, obtained by comparing ground-truth rewards from trajectory pairs.

Algorithm 1 Imitation library RLHF

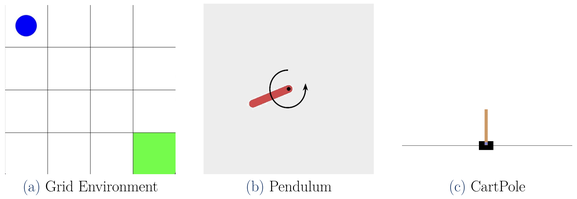
1: **Initialize** reinforcement learning (RL) model π_θ and reward model $\hat{r}_\theta$
2: **for** $i = 1, \ldots, N$ **do**
3: Rollout trajectories using π_θ
4: Generate queries from collected trajectories
5: Collect synthetic feedback for generated queries
6: Update $\hat{r}_\theta$ by minimizing the loss function on the queries and feedback
7: Train π_θ using updated $\hat{r}_\theta$
8: **end for**
9: Train a new RL model using the final learned reward model $\hat{r}_\theta$

To support new feedback types, we modified the loss function and query creation functions (steps 4-6 of Algorithm 1). We extended the synthetic feedback gatherer by adding a numeric preference strength indicator for Marginal Preference Feedback and an option for Scalar Feedback with a single trajectory and ground-truth reward.

Our adjustments can be summarized by the following steps:

1. We create an appropriate query for the selected feedback type from rolled-out trajectories.
2. We define rules to give synthetic feedback on queries (emulating human feedback) based on trajectories and their ground-truth rewards.
3. We implement an appropriate loss function for each feedback type.

We implemented a simplified RLHF setup using parts of the Imitation library with Q-learning to validate our feedback types and methods. We designed a custom Grid Environment (Figure 1a) for this evaluation. After this feasibility test, we conducted experiments in Pendulum (Figure 1b) and CartPole (Figure 1c) Gymnasium environments [6] using the Imitation library and Proximal Policy Optimization (PPO) [7].



(a) Grid Environment (b) Pendulum (c) CartPole

Figure 1: Experimental Setup: Environments

We conducted empirical evaluations using three feedback types. To ensure reproducibility, we set seeds for each evaluation. We conducted five rounds of evaluations for each environment and feedback type, using different seeds for each round, and averaged the results during training and final reward reporting. Additionally, to emulate human inaccuracy in providing Scalar Feedback, we added noise in one experiment, as this feedback type is known to be challenging for experts.

Results

All selected feedback types achieve maximum reward in the Simple Grid Environment. In contrast, a randomly initialized, untrained reward model performs poorly. We then evaluate our feedback types on Pendulum and CartPole to highlight their trade-offs. Our Pendulum findings persist in CartPole, but due to sparser rewards, the plots are less representative. Pendulum evaluations are included in this poster, while all evaluations and comparative tables are available in the paper.

Scalar Feedback with ground-truth rewards achieves high performance and low variance with even a small number of queries (Figure 2).

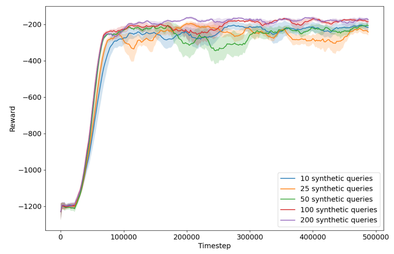


Figure 2: Ground-Truth Scalar

Given the complexity of providing precise scalar evaluations, Figure 3 shows results when the evaluator provides less accurate scalars. This degrades performance with a small number of queries, but the agent still reaches expert-level performance with 100 and 200 queries, though with higher variance for fewer queries.

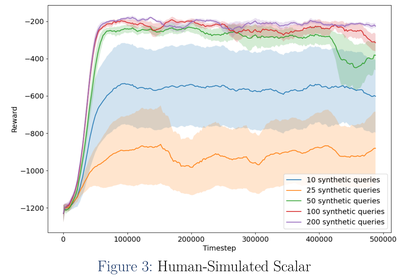


Figure 3: Human-Simulated Scalar

We confirm that while **Preference Feedback** is easier for humans to give, it conveys less information and requires more queries for good performance, as shown in Figure 4.

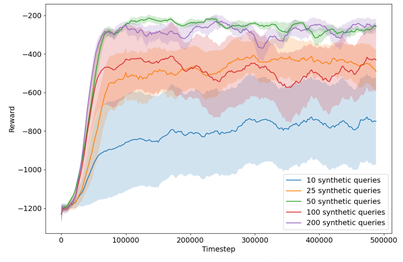


Figure 4: Preference Feedback

Results

We also report high overall variance, as was the case with human-simulated scalar feedback. However, while in Human-Simulated Scalar Feedback, more queries led to lower variance, this is not the case for Preference Feedback.

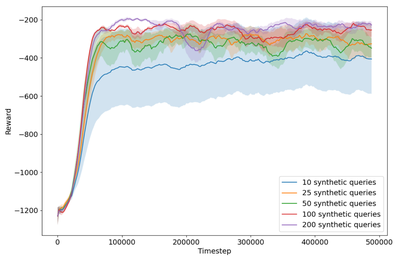


Figure 5: Marginal Preference Feedback

We observe in Figure 5 that **Marginal Preference Feedback** leads to better performance and lower variance.

Conclusions

This empirical exploration investigated the integration and effectiveness of different feedback types in RLHF. Our findings highlight the trade-off between ease of providing feedback and the amount of information conveyed [3]. These insights can assist in developing RLHF systems, offering evidence for using different feedback types based on specific needs. Exploring less common feedback types can guide future studies. Future work could incorporate real human feedback and test more complex environments to validate and expand these findings. Additionally, integrating multiple feedback types in a single agent's training presents an intriguing avenue for research.

References

[1] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, *Deep reinforcement learning from human preferences*, 2023. arXiv: 1706.03741 [stat.ML].

[2] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, *A survey of reinforcement learning from human feedback*, 2023. arXiv: 2312.14925 [cs.LG].

[3] S. Casper, X. Davies, C. Shi, *et al.*, *Open problems and fundamental limitations of reinforcement learning from human feedback*, 2023. arXiv: 2307.15217 [cs.AI].

[4] H. Touvron, L. Martin, K. Stone, *et al.*, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL].

[5] A. Gleave, M. Taufeeque, J. Rocamonde, *et al.*, *Imitation: Clean imitation learning implementations*, arXiv:2211.11972v1 [cs.LG], 2022. arXiv: 2211.11972 [cs.LG].

[6] M. Towers, J. K. Terry, A. Kwiatkowski, *et al.*, *Gymnasium*, Mar. 2023. DOI: 10.5281/zenodo.8127026. (visited on 07/08/2023).

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, 2017. arXiv: 1707.06347 [cs.LG].