Reinforcement Learning from Human Text Feedback (RLHTF)

Belen Martin-Urcelay, Andreas Krause, Giorgia Ramponi

Learning & Adaptive Systems

ETH zürich

Motivation

Reward definition is challenging

Text has more information than rankings

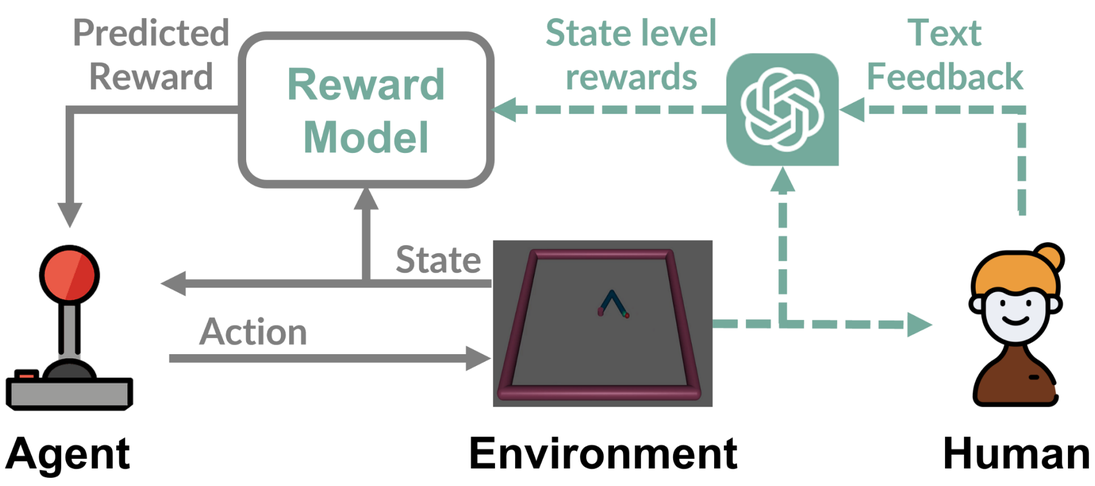
LLMs are great at processing text

Use **LLMs** as a way of harnessing the information from **human text to train** a **reward** model efficiently in **RLHF**.

Baselines

- Sentiment: We measure the sentiment of the human text feedback and apply it as a reward for all states in the trajectory [1].
- PbRL: We query human evaluators for pairwise comparisons between trajectories [2].
- True: The agent receives the true reward from the environment.

Framework



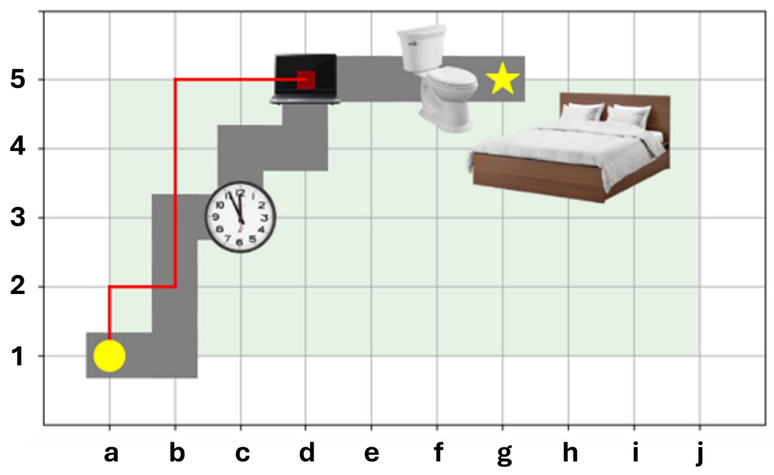
- An LLM translates human evaluations in the form of natural language into state level rewards.
- These labeled states are used to train a reward model.
- The agent is then trained with standard RL algorithms.

Algorithm

Input: number of human iterations N
Initialize Policy π_0 and reward model $\widehat{r}_0$.
for $i = 0$ **to** N **do**
  Record trajectory following policy: $\mathbf{t}_i \leftarrow \pi_i$
  Query human for feedback: $\mathbf{f}_i \leftarrow \mathbf{t}_i$
  Translate feedback to state-reward pairs with LLM: $\mathbf{s}_i, \mathbf{r}_i \leftarrow \mathbf{f}_i, \mathbf{t}_i$
  Update reward model: $\widehat{r}_{i+1} \leftarrow \widehat{r}_i, \{\mathbf{s}_t, \mathbf{r}_t\}_{t=0}^{i}$
  Update policy: $\pi_{i+1} \leftarrow \pi_i, \widehat{r}_{i+1}$
end for

Experiments in Gridworld

Goal: follow a specific path (in grey) from the start (yellow circle) to the end (yellow star). The agent's trajectory is in red.



Human Feedback
"The first move is wrong"

+

Landmark Locations
- Clock at 3c
- TV at 5d…

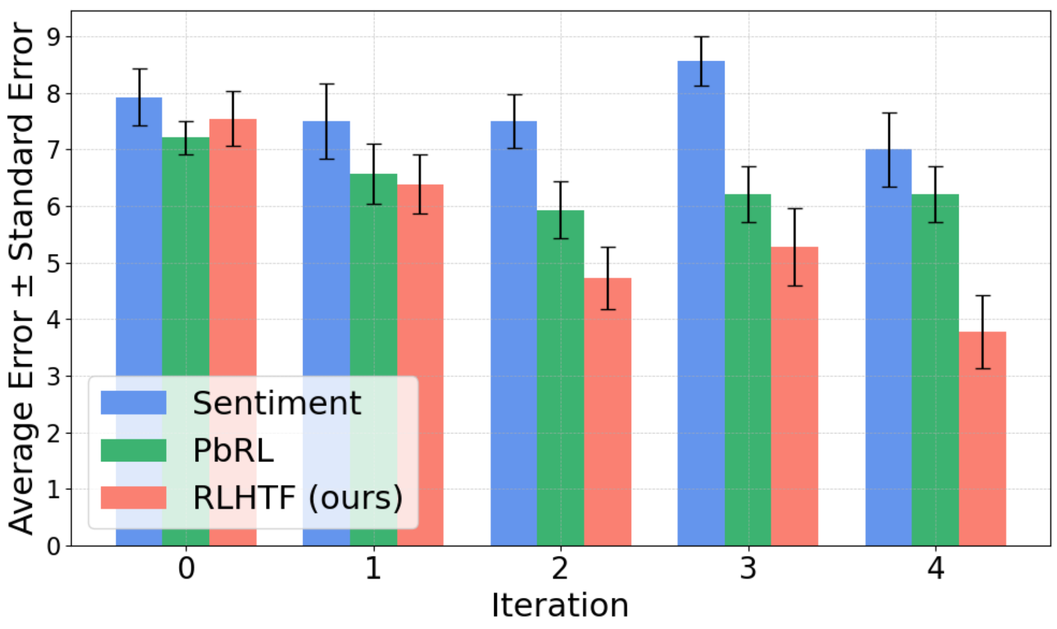
+

Trajectory
2a, 2b, 3b, 4b, 5b, 5c, 5d, 5d

→ ChatGPT →

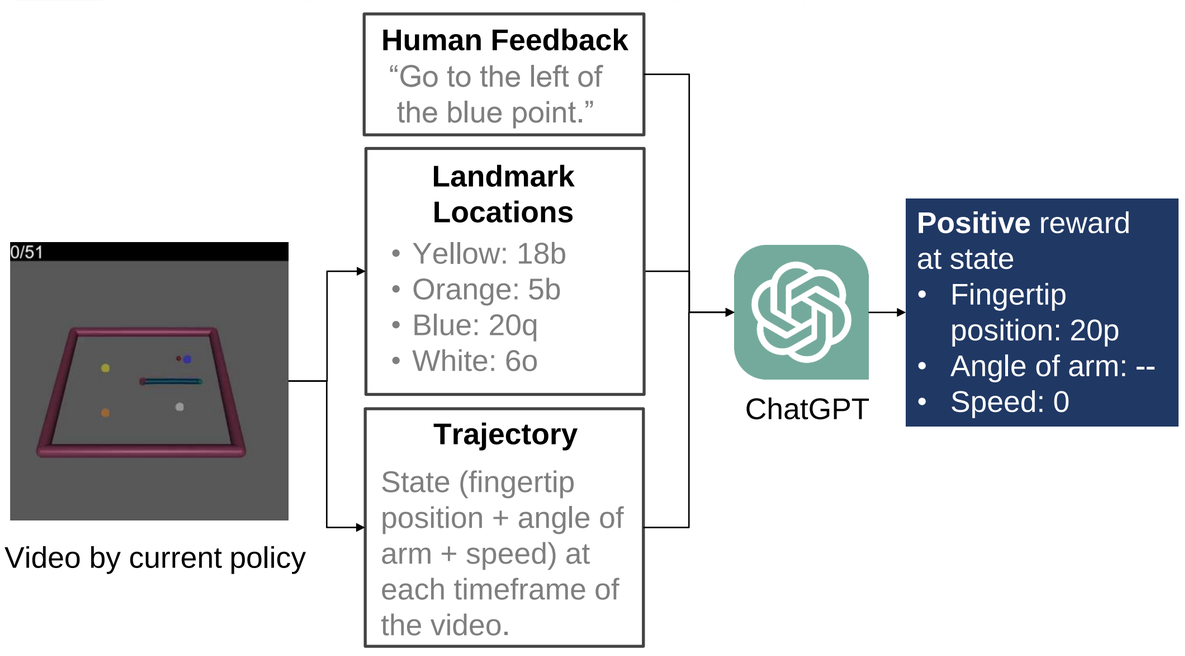
Negative reward at **2a**

Performance Comparison

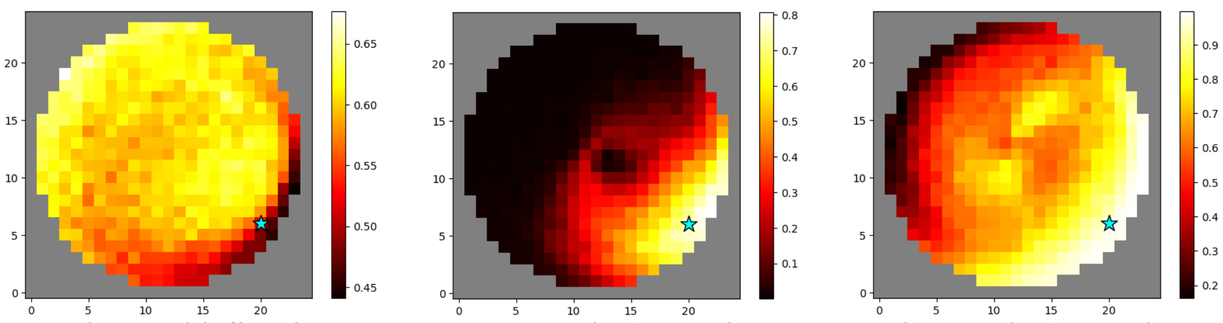


Experiments in MuJoCo

Goal: Move a two-jointed robot arm to target (red point)

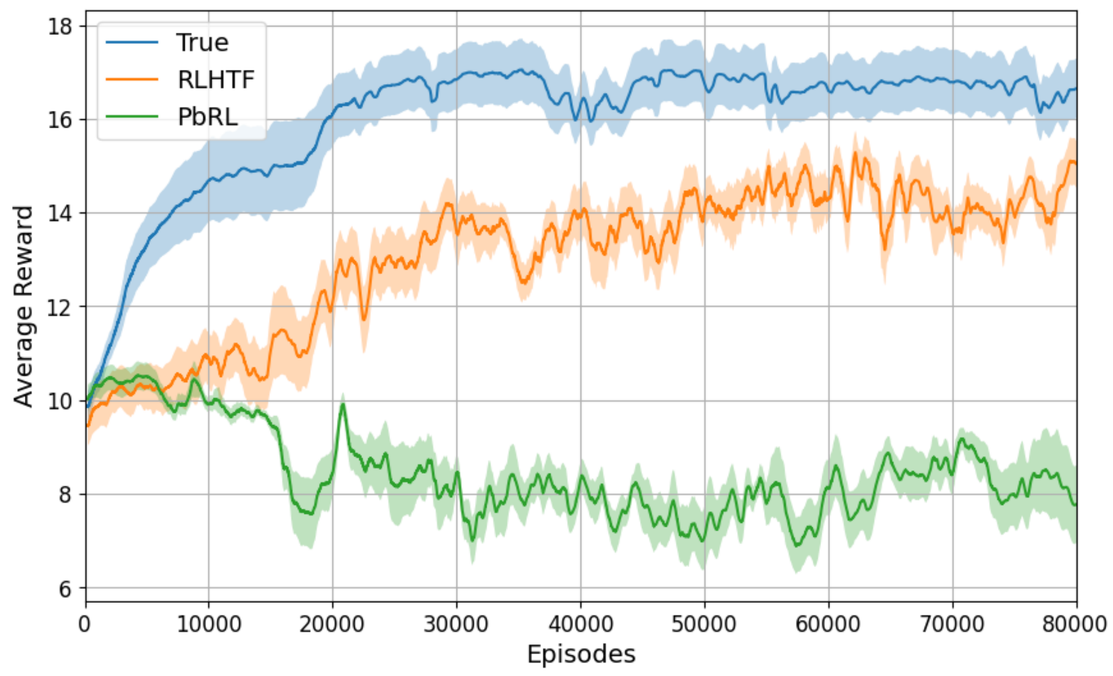


Reward Model Visualization



★: Target location □: Highest reward ■: Lowest reward

Performance Comparison



[1] Sumers, T. R., Ho, M. K., Hawkins, R. D., Narasimhan, K., and Griffiths, T. L. Learning rewards from linguistic feedback. In Proc, of Conference on Artificial Intelligence
[2] Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Proc. of Advances in Neural In formation Processing Systems.