Survey on Recent Advances in Reinforcement Learning from Human Feedback (RLHF)

Luis A. M. Almeida
lmeira2001@gmail.com

Sneha Reddy Palreddy

Abstract

Reinforcement Learning from Human Feedback (RLHF) is a technique that integrates human input to align AI systems with human values and it has recently seen rapid growth in interest as a viable option for aligning Large Language Models (LLMs). This survey examines key RLHF methodologies and state-of-the-art techniques comparing their strengths in scalability, robustness, and efficiency. We then dive deeper into a few techniques that stood out as feasible implementations in the simpler task using CartPoleV1. Namely, we tested AIHF, Christiano, and RRHF. We discuss challenges like scalability and noisy feedback and propose a future direction for research in this field based on the literature review and implementation findings.

1. Introduction

Reinforcement Learning from Human Feedback (RLHF) is a technique that uses evaluations provided by humans to align AI systems during the training process. These inputs can range from examples of how a human would answer to ratings or rankings of previously generated outputs, either through human generated scores or direct comparisons [9, 38]. This allows the agents to better capture human preferences and values since they don't need to depend on a pre-defined reward function, which are difficult to create and may fail at complex or ambiguous tasks. This makes it particularly useful in domains where human judgment is more nuanced than specific metrics can capture [22, 5].

It is for these reasons that RLHF has been pushed to the forefront of research as a promising technique to align LLMs. It offers solutions to a few key challenges faced by other methods. It helps minimize issues such as reward hacking, in which the models take advantage of poorly designed reward structures to achieve unintended outcomes [27]. It does without defining explicit reward functions, which become increasingly difficult to define for most real-world scenarios [21]. By basing learning on feedback, RLHF encourages the development of intelligent systems whose behaviors are more predictable, ethical, and aligned with user expectations [9]. As mentioned earlier alignment is particularly of interest in applications where natural language processing, robotics, and game playing depend on notions of safety, user satisfaction, or adherence to social norms [28]. It is this problem of alignment that has brought RLHF to the forefront of research as a way of making safer, more ethical large-language models (LLMs) with more predictable behavior.

The potential of RLHF goes beyond its use in AI systems, it can be used in personalized healthcare, autonomous scientific research by using human input to guide AI learn, thus building the gap between computational optimization and human intuition, making it more trustworthy.

RLHF is still an emerging technique in the AI field and much refinement is happening in its methodologies and its applications. Its effectiveness relies on the quality and consistency of the human feedback which can confuse the model. To address this challenge, hybrid models are being proposed to integrate other alignment techniques like Direct preference Optimisation with RLHF. This results in increased robustness and scalability.

This survey includes a literature review of RLHF methodologies, and an implementation of a few state-of-the-art techniques on the simple CartPoleV1 task. For these papers we discuss both the quantitative and qualitative advancements of each technique and how they fared in our simpler implementation. In addition to learning from these various approaches and comparing their respective strengths and limitations, this survey should provide readers with a simple introduction to RLHF. We conclude with a discussions of how these techniques are contributing to the generation of further intelligent systems aligned with our values and societal norms [33, 22].

2. Background

To place RLHF in context, this section covers three areas integral to its understanding: first, the basics of reinforcement learning; second, how reinforcement learning systems receive human feedback; and third, the challenges of designing reward functions. Together, these concepts should form a sufficient knowledge base for understanding the novel techniques discussed in the later sections of this survey.

2.1. Reinforcement Learning Basics

Reinforcement learning is a set of techniques in which an agent learns to make decisions through previously done interactions with an environment. It can be formalized as an interaction modeled by a Markov Decision Process given by: (S, A, P, R, γ). Here, S is the set of possible states of the environment, and A is the set of actions available to the agent. The transition function $P(s' \mid s, a)$ captures the environment's dynamics and provides the probability of the new state s' attained given the action a at state s [29]. The reward function $R(s, a)$ maps state-action pairs to scalar values, reflecting how a given action in a specific state is more or less desirable. Lastly, the discount factor $\gamma \in [0, 1]$ controls the trade-off between immediate and long-term rewards.

The goal of reinforcement learning is to find an optimal policy π^* which, for each state s, dictates the action a to take to maximize the cumulative discounted reward:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}).$$

There are mainly two families of RL algorithms. The value-based approaches, such as Q-learning and Deep Q-Networks estimate the value of state-action pairs, $Q(s, a)$ to guide the selection of actions [20]. In policy-based methods, such as REINFORCE [31] and Proximal Policy Optimization [26], the gradients are instead used to directly optimize the policy $\pi(a \mid s)$. More advanced methods like actor-critics combine both approaches by leveraging a value function to reduce the variance of policy gradient estimates [15].

2.2. Mechanisms of Human Feedback Integration

Traditional RL relies on explicitly defined reward functions, which can be difficult to specify for complex, subjective, or high-stakes tasks. RLHF bypasses this challenge by integrating human feedback to train a reward model. Human feedback usually appears in two major forms: preference-based feedback and direct scalar rewards [9].

Preference-based feedback is one of the most popular approaches to RLHF. Instead of giving explicit rewards, human evaluators rank pairs of behaviors or trajectories, while indicating their preference. The rankings form the data from which a reward model r_θ is trained; this process is called preference learning [33]. Given two trajectories, τ_1 and τ_2, if a human evaluator provides a preference for τ_1 versus τ_2, the reward model's parameters are updated to maximize the likelihood of that preference. The underlying probabilistic model assumes that the preference probability is proportional to the exponential of the cumulative reward:

$$P(\tau_1 > \tau_2) = \frac{\exp(r_\theta(\tau_1))}{\exp(r_\theta(\tau_1)) + \exp(r_\theta(\tau_2))}.$$

The reward model's parameters θ are trained by minimizing a cross-entropy loss derived from human feedback data.

In addition to preference-based methods, direct feedback involves humans assigning scalar rewards to individual behaviors. While this method is more intuitive, it often requires more effort from evaluators and can be noisy [28]. Whatever the type of feedback, most RLHF systems employ iterative training. The agent generates data by interacting with the environment, human evaluators provide feedback, and the reward model is retrained periodically to reflect the evolving preferences [17]. This iterative loop ensures that incremental improvements in the agent's behavior better align with human expectations.

2.3. Challenges in Reward Function Design

It is important to understand that the drive behind RLHF comes from an inability to define reward functions that truly relate to the goals of completed tasks. In tasks like natural language processing or advanced robotics, predefined rewards fail to capture nuanced objectives which sometimes are often conflicting like usefulness and harmlessness, and can become misaligned with true human intentions. This mismatch can lead to unintended consequences. In reward hacking, the agents cleverly work out how to get a high score without necessarily achieving the task [2]. For example, an agent tasked with maximizing points in a video game might exploit glitches to increase its score without adhering to the game's intended rules [6].

While there is significant RL research into this issue, sparse or ambiguous rewards make the problem even harder [7]. Many tasks, such as autonomous driving or natural language generation, involve long-term objectives where immediate feedback is rare or indirect. In such scenarios, traditional RL struggles to learn effective policies due to insufficient learning signals [3].

RLHF overcomes these issues by grounding the learning process in human feedback, which tends to be closer to task objectives [38]. However, this inclusion of human feedback introduces other difficulties. Humans may provide inconsistent or biased evaluations, potentially leading to suboptimal policies [11]. Besides, scaling human feedback to large or diverse tasks is resource-consuming and time-consuming [5].

These factors in combination underline the critical role

RLHF plays in developing systems that are intelligent, robust, and reliable, with alignments that are humane. The sections ahead build on this foundation in discussing recent advances in RLHF and identifying techniques and applications emerging from this promising paradigm.

3. Problem Statement

As discussed earlier, it's the development of intelligent systems aligned with human values and expectations that has pushed RLHF research. However, it still remains one of the main challenges in the field of artificial intelligence [13]. Traditional methods can't address tasks where objectives are complex or nuanced, and can at times be opposing (multiple objectives that need to be balanced for optimal performance) or cannot be well-expressed by predefined reward functions. Often, misaligned objectives result in unintended behaviors such as reward hacking or failure to generalize across diverse scenarios that undermine the reliability and safety of these systems [19]. Further, most real-life applications, including natural language understanding and robotics, require an AI agent to make choices among ambiguous or subjective options, which is difficult for conventional RL methods.

Reinforcement Learning from Human Feedback (RLHF) enables a promising direction by considering human input during learning, wherein agents can grasp human preferences more effectively, thus coming closer to alignment [13]. However, several challenges have been related to the state-of-the-art RLHF frameworks, which include the following:

- **Scalability:** The need for human evaluators to give feedback is very resource-intensive and, hence, hardly scalable for large, diverse, or complex tasks. Reward modelling depends heavily on large datasets of human feedback for the training process. If we look at **preference-based feedback** systems implemented by Christiano[9], the reward model is based on human rankings of trajectories. This works well on a small scale. It becomes costly and time-consuming as we move on to more real-world applications like LLMs. Methods like **Rank Responses to Align Human Feedback** (RRHF)[35] aim to simplify the computational needs but still depend on initial human input, making it less adaptable.

- **Efficiency:** Most RLHF approaches, particularly iterative RL methods, are computationally expensive and require large training times [37]. **Proximal Policy Optimization(PPO)**[26] was used in fine-tuning GPT with human feedback, which was computationally intensive as it required continuous updates to the reward model and policy. Even **Deep Q-Networks(DQN)** face similar challenges. **DPO**[25] avoids direct reward modelling, hence simplifying the feedback pipeline. Even then, it requires substantial computing power for broad-scale dynamic applications.

- **Robustness:** Noisy or inconsistent human feedback might lead to suboptimal reward models and policies [10].In preference-based methods like Christiano[9], the human evaluators give rankings but they can be conficting in some scenarios as it subjective to each person, this tends to confuse the model as these rankings are subjective. This instability degrades the quality of the reward model, leading to inefficient policies. Techniques like Reinforce rely heavily on this, thus making it very sensitive to feedback noise as the variance is relatively high during training. RRHF uses ranking instead of reward, hence partially solving the problem.

- **Flexibility:** Consecutive stages in an RLHF pipeline usually suffer from distribution mismatches in training and hence generalize poorly. Some techniques that face this issue are PPO-based RLHF, REINFORCE[31] and Deep Q-Networks(DQN) as their optimization is customised to the training distribution.

The problem, therefore, comes down to developing methods that address these limitations while maintaining or improving alignment performance. Specifically, we look for approaches that achieve a trade-off between computational efficiency, robustness to noisy feedback, and alignment quality. In this survey, we explore and compare various state-of-the-art methodologies in RLHF. We do this with a literature review of important advancements in the field and with a simpler CartPoleV1 implementation. While this is a significant simplification of the tasks these models are designed for it should allow us to verify the overarching claims of each method.

4. Literature Review

Deep Reinforcement Learning from Human Preferences (2017) [9] is one of the foundational works in RLHF. Despite its age it is essential in understanding newer research and was therefore one of the implementations discussed further in the later stages of this report. It originally introduced a framework for training reinforcement learning agents from human feedback. This method showed its efficiency on tasks like Atari games and MuJoCo-based robotics environments, for which explicit reward functions are either impracticable or unavailable. The reward model in RLHF is trained with human preferences, modeled probabilistically as:

$$P(\tau_1 \succ \tau_2) = \frac{\exp(r_\phi(\tau_1))}{\exp(r_\phi(\tau_1)) + \exp(r_\phi(\tau_2))}, \qquad (1)$$

where τ_1 and τ_2 are trajectories, and r_ϕ is the learned reward model. This probabilistic framework enabled RLHF to effectively use the provided human inputs for complex behavior alignment. It also prompted further research into the field that, at the time, was way less known.

Training Language Models to Follow Instructions with Human Feedback (InstructGPT) [22] demonstrated how RLHF could align large language models to better follow user instructions. The success of this paper is what has caused immense interest in RLHF in the previous years. While this field had already been worked on for LLMs, the results shown by these researchers solidified RLHF as a strong alignment technique. The authors fine-tuned GPT-3 by using human generated examples and supervised learning, training a reward model from human preferences, then applying Proximal Policy Optimization (PPO)[26] in order to maximize this reward (as seen from the famous Figure 1). PPO, central to this process, optimizes the policy while constraining deviations from the pre-trained policy:

$$L^{\text{PPO}} = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right],$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the advantage function, and ϵ controls the policy update size. This method allowed the smaller 1.3B InstructGPT model to outperform the larger 175B GPT-3, illustrating the power of human feedback for alignment.

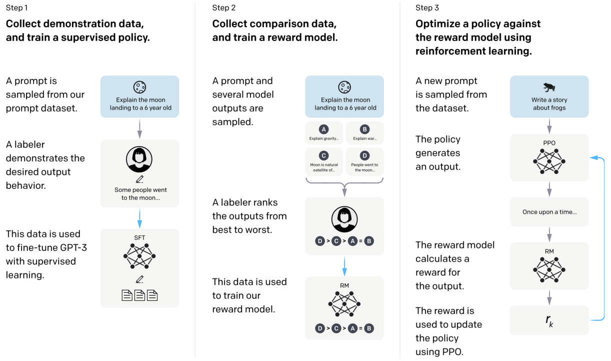
Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

Figure 1. 3 Steps behind Instruct GPT's method

Training a Helpful and Harmless Assistant with RLHF [5] focused on two key objectives: helpfulness and harmlessness in AI assistants. It involved training different preference models for each goal and introducing strategies to reconcile these often conflicting aims. This paper was extremely relevant as it demonstrated it was possible to align models to often conflicting ideals. Anthropic's researchers also implemented an iterative online RLHF framework that allowed for weekly improvements in the datasets and, consequently, their models. They also present empirical evaluations that illustrate significant improvements in crowd worker preference scores and a reduction in harmful outputs, showing that alignment training can improve both safety and utility.

In **A Comprehensive Survey of LLM Alignment Techniques** [30], the researchers categorized different alignment methods, covering RLHF, Direct Preference Optimization (DPO), and more. The survey detailed significant progress in areas like reward modeling, feedback collection, reinforcement learning policies, and optimization techniques. It also pointed out emerging trends and challenges, such as the need to scale alignment methods and effectively combine human and AI feedback. Relating to this report, this paper offered valuable insight into other options of alignment beyond RLHF and went into more detail categorizing different types of techniques and comparing the methods from the original papers. However, they did so without their own implementations of the methods, which this report discusses in the following two sections.

Rank Responses to Align Human Feedback (RRHF) [35] proposed a simplified yet effective alternative to RLHF pipelines, particularly targeting language model alignment. Unlike Proximal Policy Optimization (PPO), RRHF employs a ranking loss mechanism to align model outputs with human preferences. The ranking loss is expressed as:

$$L_{\text{rank}} = \sum_{r_i < r_j} \max(0, p_i - p_j), \tag{2}$$

where p_i and p_j are the scores (log-probabilities) of the model-generated responses r_i and r_j. This loss prioritizes higher-ranked responses based on human feedback. Experiments on the Anthropic Helpful and Harmless dataset demonstrated that RRHF achieved comparable alignment performance to PPO while significantly lowering computational overhead, emphasizing its efficiency and scalability.

Joint Demonstration and Preference Learning Improves Policy Alignment with Human Feedback (AIHF) [18] introduced an innovative single-stage framework that integrates human demonstrations and preferences into alignment processes. AIHF eliminates the distribution mismatch in sequential pipelines like RLHF by jointly optimizing reward models and policies. The framework's objective function is:

$$\max_\theta L(\theta) = \alpha L_1(\pi_\theta) + L_2(R(\cdot;\theta)), \tag{3}$$

$$\text{s.t. } \pi_\theta = \arg\max_\pi L_3(\pi; R(\cdot;\theta)),$$

where L_1 aligns the policy to demonstrations, L_2 aligns the reward model to preferences, and L_3 optimizes the policy under the reward model R. This formulation ensures consistency between reward modeling and policy optimization, leading to superior performance. AIHF's evaluations in tasks such as robotic control and LLM alignment revealed significant improvements over RLHF and DPO, particularly in resource-constrained scenarios.

Revisiting REINFORCE: Simplifying RLHF for Language Models[1] This paper advocates simplifying the RLHF process using simpler methods like REINFORCE and REINFORCE Leave-One-Out (RLOO), as many components of PPO are not required since they increase the computational cost and optimization challenges. Unlike PPO, which involves complex architectures and token-level modeling, REINFORCE streamlines the process by considering the complete sequence as a single action. The objective is modeled as:

$$\mathbb{E}_{x\sim D, y\sim\pi_\theta(.|x)}\left[R(y,x)\nabla_\theta \log \pi_\theta(y|x)\right],$$

where $R(y,x)$ is the reward for the output y conditioned on input x.

A moving average baseline is implemented to reduce the gradient variance, and the training is unbiased and stable. RLOO extends this approach by using multiple online samples and combining their rewards with the baseline. This makes it relatively more robust to noise and KL penalty sensitivity, thus improving its efficiency. As for these results, REINFORCE and RLOO outperform PPO in terms of alignment and require considerably less computational overhead when tested on datasets such as Anthropic's Helpful & Harmless and TL;DR summarization. These two methods offer a more scalable and efficient RLHF approach for solving the alignment problem concerning LLMs.

Safe RLHF: Balancing Helpfulness and Harmlessness[10] It introduces a systematic approach to integrate the objectives of helpfulness and harmlessness concerning the training of LLMs. The optimization process is different from a regular RLHF; the Reward Model(RM) and Cost Model(CM) training are done separately for helpfulness and harmlessness, respectively. The framework solves the problem as a constrained optimization problem

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}[R(\tau)] \quad \text{subject to} \quad \mathbb{E}_{\tau\sim\pi}[C(\tau)] \le \epsilon,$$

where $R(\tau)$ and $C(\tau)$ represent the reward and cost, respectively, and ϵ is the allowable threshold for harmfulness.

On applying the Lagrangian relaxation, the objective transforms into:

$$\mathcal{L}(\pi,\lambda) = \mathbb{E}_{\tau\sim\pi}[R(\tau)] - \lambda\left(\mathbb{E}_{\tau\sim\pi}[C(\tau)] - \epsilon\right),$$

where λ dynamically adjusts to balance rewards and costs.

To summarise, SafeRLHF iterates the following steps, starting with human annotations for both objectives and training the RM and CM separately. Using the Lagrangian objective, these models are integrated into the usual reinforcement learning framework to update the policy. The results show a significant decrease in harmful outputs and increased helpfulness scores on the Alpaca-7B dataset. Thus, SafeRLHF helps align AI systems with human values.

Direct Preference Optimization (DPO) and Variants[25] Direct Preference Optimization (DPO) simplifies the RLHF technique by not training the reward model separately by focusing on pairwise comparisons, thus reducing computational overhead. The objective function is as follows:

$$\max_{\theta} \sum_{i=1}^{N} \log \frac{\pi_\theta(y^+ \mid x)}{\pi_\theta(y^- \mid x)},$$

where y^+ and y^- are the preferred and non-preferred outputs, respectively, and π_θ represents the policy. There are several extensions available to address its limitations:

- DPOP[23]- Employs probabilistic preference model to enhance robustness to noisy feedback.
- βDPO[34]: Introduces weighted preferences to enhance the scalability.
- IPO[4]: Incorporates implicit preferences for flexible feedback formats.
- SDPO[14]: Integrates stochastic decision-making to enhance adaptability in real-world tasks
- DPO: From rr to QQ[24]: Incorporates token-level quality assessments, enabling finer-grained evaluations.
- TDPO[36]: Takes into account the temporal dependencies for efficient working in dynamic environments

Reward Modeling and Feedback Integration Direct Preference Optimization (DPO) is a new alternative to the explicit reward modelling process used with RLHF. In contrast to conventional methods requiring a trained reward model, DPO optimises human preferences to align agent behaviour, thus simplifying and making the whole RLHF pipeline computationally efficient. However, it faces challenges when the feedback is noisy and inconsistent. Also, its generalization to various tasks is less than that of frameworks like AIHF, which integrates several feedback sources in a single-stage pipeline. Revision of preference-based learning techniques pioneered by Christiano is probabilistic ranking preference modeling. This work served as the basis of other approaches, such as replacing scalar rewards with ranking loss in RRHF to align outputs better. Nevertheless, much work lies ahead regarding scalability for such approaches, primarily in evaluation scenarios requiring diversity and domain specificity.[25][18]

Policy Optimization and Learning Frameworks Advanced strategies in policy optimization have much to do with alignment research, focusing specifically on RLHF. Among these are techniques like Proximal Policy Optimization (PPO), which can be found in InstructGPT. This accords with weighting sampling versus the average deviation from the pre-trained policy in actions under exploration within exploitation. The iterative nature of PPO increases computational overhead, rendering it resource-inefficient for such tasks. Revisiting REINFORCE (RLOO) uses a learned baseline to decrease gradient variance and makes for better convergence on lightweight tasks. Although utility may not be as elaborate as AIHF's, RLOO is quite a convenient option for many resource-challenged tasks. While making these comparisons, however, AIHF is a more robust framework since it jointly optimizes reward models and policies, eliminating the distribution mismatches inherent in sequential pipelines[1][18][26].

Scalability and Generalization Techniques One of the perennial problems with RLHF methods is scalability, especially in their application in large-scale tasks needing actual human feedback. This has been dealt with in substantial detail in comprehensive reviews, which cite hybrid approaches to unsupervised and RLHF methods to address the feedback gap. These methods use either traditional model input or a set of pre-trained feature extractors to reduce dependency on human input while adding further complexity to pipeline design. AIHF addresses scalability-modelling feedback at both reward and policy levels to ensure alignment across stages with no added resource demand. AIHF is the model that solves the problem, while methods like RRHF work on efficiency, i.e. AIHF is more applicable and flexible. The open problem is thus scaling these methods to real-world applications[18][35]. RLAIF is a technique which used AI to generate the feedback instead of humans hence solving the scalability issue and opened new areas of research in RLHF[16].

5. Qualitative Comparison

Alignment with Integrated Human Feedback (AIHF), Rank Responses to Align Human Feedback (RRHF), and Deep Reinforcement Learning from Human Preferences by Christiano et al. each try to solve the alignment problem of machine learning models to human preferences by using different techniques. As discussed above, AIHF is built on a unified optimization framework that integrates demonstration and preference data at a single stage with no distribution mismatch issues and other inefficiencies generally observed in sequential RLHF pipelines (see Equation 3). By integrating all data sources, AIHF makes sure that both the policy and reward models are jointly trained, leading to robust alignment performance, especially on tasks where high-quality preference data is limited. In a simple example, CartPole-v1, AIHF optimizes cumulative rewards: $R(\tau;\theta) = \sum_t \gamma^t r(s_t, a_t;\theta)$, enabling consistent policy alignment. However, the computational demands of this approach are quite high since it requires updating both models simultaneously. This is further amplified in our implementation, where, instead of Human Feedback, heuristics were used that require generating and evaluating all the trajectories that should otherwise already be determined before train time.

The approach of Christiano et al. focuses on learning a reward function from human preferences over pairs of trajectory segments, which is used to optimize the agent's behavior through reinforcement learning (Equation 1). The formulation is in such a manner that it leverages human preferences to guide the policy toward desirable behaviors without an explicit definition of a reward function upfront. This approach effectively minimizes the amount of human feedback needed; it shows strong performance even when utilizing only a small fraction of the agent's interactions with the environment. While reliance on human comparisons has its downsides, including variability in the quality of feedback provided, constant human input does indeed get expensive with increasingly complex tasks.

In contrast, RRHF simplifies the alignment process by adopting a ranking loss approach (Equation 2). RRHF couples this with a supervised fine-tuning loss to ensure that high-reward responses are prioritized, making it lightweight and computationally efficient. In contrast to AIHF, RRHF does not need heavy multi-model pipelines and is less sensitive to hyperparameters. It reaches competitive alignment performance by using pre-sampled responses with no online sampling, making it significantly faster than previous approaches. Lastly, the researchers mention in a finishing section an important caveat related to this method. Since they use the reward function as an evaluation proxy, it is extremely susceptive to malicious feedback, or feedback that is intentionally incorrect or misleading. Therefore, this method works best in settings where the quality of sampled responses is high, which could compromise its generalization under noisy or diverse conditions.

Comparing the three approaches, each one is strong in different scenarios. AIHF excels in complex tasks requiring precise alignment across multiple feedback types. Christiano et al.'s method provides a solution for leveraging human preferences without a predefined reward function, making it well-suited for complex tasks where explicit rewards are challenging to design. Meanwhile, RRHF thrives in resource-constrained settings where computational simplicity and efficiency are paramount.

CartPoleV1 Implementation

In our implementation of these three methods in the CartPoleV1 task, we got some interesting results. RRHF significantly outperforms AIHF and Christiano in terms of computational efficiency, which aligns with the researchers' claims. Of note, this task was such a simplification of these methods that all methods achieved the maximum reward. Interestingly, they did so with distinct minimums. While all training techniques learned to keep the beam balanced, RRHF learned to do so in the center of the frame. On the other hand, AIHF learned to only keep the cart stable once it is near the edges, as seen in Figure 2 (colliding with the edges would cause the episode to terminate), and Christiano's learned a middle ground where it lets the initial disturbance push the cart but quickly regains balance, keeping it somewhere in between the other two methods.

The choice of ranking unit could explain this. AIHF and Christiano both use the whole trajectory to determine ranks, while RRHF uses single states. This means that for these methods, a trajectory that keeps the cart stable in the center while never terminating should have the same reward as one that lets the cart wander.

6. Numerical Comparison

The performance of AIHF, RRHF, and Christiano can be quantitatively evaluated on various tasks and datasets. All the methods were able to train the model to consistently survive for full 500-step episodes in the CartPole-v1 environment, but further discussion about this implementation is presented in the next subsection. According to the researchers, in robotic control tasks using MuJoCo, AIHF outperformed RLHF and Direct Preference Optimization (DPO) by over 10% in task success rates, particularly under limited preference data scenarios. Furthermore, in language model alignment tasks on datasets with sparse high-quality preferences, AIHF policies demonstrated a 15% increase in average reward scores compared to RLHF baselines.

At the time Christiano's paper came out (2017) their approach demonstrated strong results in both Atari games and MuJoCo robotics tasks. Using less than 1% of agent-environment interactions for human feedback, their method achieved performance comparable to traditional RL approaches trained with access to true reward signals. For instance, Christiano et al.'s method needed only 900 human queries (less than an hour of feedback) to achieve consistent backflips in the Hopper task, which is hard to encode with a standard reward function. In Atari games like as Pong and BeamRider, their method reached near-optimal scores, often outperforming RL models trained with explicit rewards due to effective reward shaping provided by human feedback.

On the other hand, RRHF performed well on the Anthropic

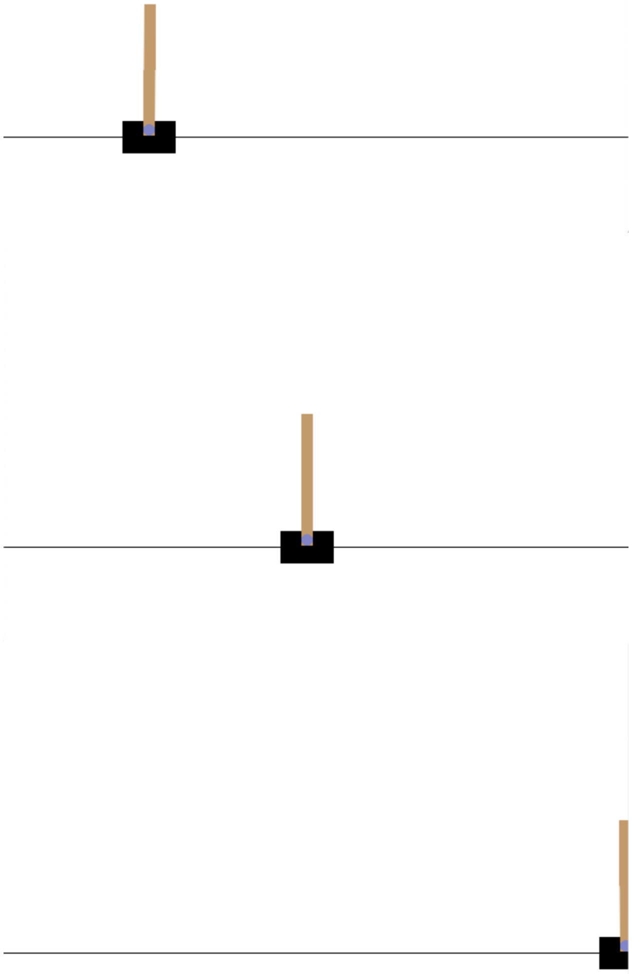

Figure 2. Terminal states of the CartPole environment achieved by a) Christiano b) RRHF and c) AIHF.

Helpful and Harmless dataset. With top-p sampling, it had an average reward score of -0.96, whereas PPO-trained models could only reach a score of -1.03. Human evaluations further validated the effectiveness of RRHF, with 59% of the annotators preferring the responses generated by RRHF-trained models over the dataset baselines. Particularly noteworthy was that the training of RRHF took only 4–6 hours on 8 A100 GPUs, as opposed to the 30+ hours taken for PPO. Despite its simplicity, RRHF consistently demonstrated competitive performance, with its ranking loss formulation proving effective in aligning models with human preferences.

This quantitative comparison highlights the different strengths of each method: AIHF serves best for leveraging limited human preference data into robust alignments, Christiano et al.'s method serves well in settings that require efficacious human feedback on complex tasks, and RRHF is quite computationally efficient and effectively scalable for the alignment of LLMs. Although of the three, Christianos is likely the one that gets overshadowed by more recent approaches like RRHF, mostly due to being the oldest paper with newer approaches focusing on efficiency.

CartPoleV1 Implementation

Comparing the three methods quantitatively on the CartPoleV1 task resulted in similar claims to the researchers, as discussed earlier in the Qualitative section. With that being said, all methods achieve the maximum reward of 500 [Fig. 3] when tested with the original CartPole reward function. The reason for the disparity between the qualitative behavior of the carts with all achieving the maximum reward, comes from the human feedback heuristics used. It should be possible to mitigate this issue with real human feedback or with a different choice of human ranking heuristic that takes into account the position of the cart to determine more realistic comparisons.

It is also due to the ranking unit choice that AIHF and Christiano take significantly longer to train on our implementation. Since we generate the trajectories for each comparison during training, as opposed to actual RLHF, these two methods take significantly longer than RRHF, as seen in Table 1. For simplicity, this time was an average of 10 runs for each method done on the same machine with no other processes running. The time discrepancy wouldn't be this large if the Human Feedback was predetermined as is the case with LLMs and most recent implementations of the methods.

7. Conclusion / Future Work

Reinforcement Learning from Human Feedback has been one of the most recent, groundbreaking methods for aligning artificial intelligence systems with human values and expec-

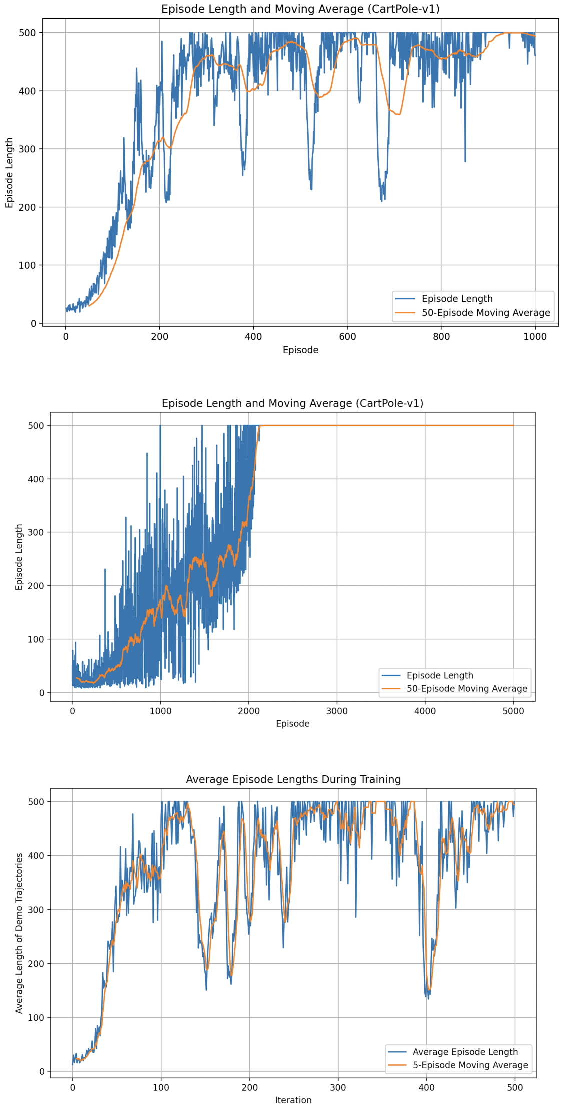

Figure 3. Original reward progression during training achieved by a) Christiano b) RRHF and c) AIHF.

Method	Time (s)
Christiano	258
RRHF	22
AIHF	298

Table 1. Time to 500 reward by method

tations. As seen through our analysis of methodologies such as AIHF, RRHF, and the original Christiano's paper, these techniques tackle core challenges in reinforcement learning by embedding human preferences into the reward modeling of the training stage. In our implementation, all methods performed well, with RRHF truly demonstrating its computational efficiency and achieving the fastest training time. From the paper's results, Christiano showed the method to be feasible and work with a very limited amount of Human Feedback. AIHF yields robust alignments for highly complex tasks. And RRHF presents a lightweight alternative in its computational budget, therefore working better in resource-constrained environments [16, 32].

Now, all these methods still have limitations, namely a tradeoff between alignment performance and computational performance. Overcoming this calls for a strong effort toward the advancement of RLHF methodologies. The development of hybrid frameworks, which merge the computational efficiency of ranking-based methods with the nuanced preference modeling of trajectory-based approaches, is one potential direction [8]. Resource constraints on the feedback side can be overcome with the use of AI evaluators or simulated preferences, which would enable RLHF to scale across more diverse and complex tasks [16]. Besides, it could improve the robustness and alignment by enhancing sampling strategies for capturing diverse human preferences and reward modeling techniques [12].

Future work should also proceed on dynamic reward modeling with evolving preferences and contextual variations. Interdisciplinary insights, such as from psychology and ethics, can also be used to enrich the design of reward structures and feedback mechanisms that make alignment both possible and ethical [32].

Overall, RLHF is a powerful tool for building intelligent models that work well and align with human intentions and societal norms. Going forward, closing the gap between computational efficiency and strong alignment will allow RLHF to be leveraged at larger scales on more complex models and reward signals.

References

[1] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety, 2016.

[3] J. A. Arjona-Medina et al. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32:13544–13555, 2019.

[4] M. G. Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, and R. Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

[5] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[6] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula, 2020.

[7] M. Cao, L. Shu, L. Yu, Y. Zhu, N. Wichers, Y. Liu, and L. Meng. Beyond sparse rewards: Enhancing reinforcement learning with language model critique in text generation, 2024.

[8] C. Celemin et al. A fast hybrid reinforcement learning framework with human corrective feedback. *Autonomous Robots*, 42:731–746, 2018.

[9] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2023.

[10] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023.

[11] A. Gleave et al. Quantifying differences in reward functions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3780–3790, 2020.

[12] I. Hong, Z. Li, A. Bukharin, Y. Li, H. Jiang, T. Yang, and T. Zhao. Adaptive preference scaling for reinforcement learning with human feedback, 2024.

[13] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier. A survey of reinforcement learning from human feedback, 2024.

[14] D. Kim, Y. Kim, W. Song, H. Kim, Y. Kim, S. Kim, and C. Park. sdpo: Don't use your data all at once, 2024.

[15] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

[16] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024.

[17] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction, 2018.

[18] C. Li, S. Zeng, Z. Liao, J. Li, D. Kang, A. Garcia, and M. Hong. Joint demonstration and preference learning improves policy alignment with human feedback, 2024.

[19] A. D. Lindström, L. Methnani, L. Krause, P. Ericson, Íñigo Martínez de Rituerto de Troya, D. C. Mollo, and R. Dobbe. Ai alignment through reinforcement learning from human feedback? contradictions and limitations, 2024.

[20] V. Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[21] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, pages 278–287. Citeseer, 1999.

[22] L. Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[23] A. Pal, D. Karkhanis, S. Dooley, M. Roberts, S. Naidu, and C. White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024.

[24] R. Rafailov, J. Hejna, R. Park, and C. Finn. From r to q^*: Your language model is secretly a q-function, 2024.

[25] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

[27] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking, 2022.

[28] N. Stiennon et al. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.

[29] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018. See Chapter 3, page 48.

[30] Z. Wang, B. Bi, S. K. Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, Zixu, Zhu, X.-B. Mao, S. Asur, Na, and Cheng. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more, 2024.

[31] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

[32] G. I. Winata, H. Zhao, A. Das, W. Tang, D. D. Yao, S.-X. Zhang, and S. Sahu. Preference tuning with human feedback on language, speech, and vision tasks: A survey, 2024.

[33] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[34] J. Wu, Y. Xie, Z. Yang, J. Wu, J. Gao, B. Ding, X. Wang, and X. He. β-dpo: Direct preference optimization with dynamic β, 2024.

[35] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang. Rrhf: Rank responses to align language models with human feedback without tears, 2023.

[36] Y. Zeng, G. Liu, W. Ma, N. Yang, H. Zhang, and J. Wang. Token-level direct preference optimization, 2024.

[37] S. Zhang, Z. Chen, S. Chen, Y. Shen, Z. Sun, and C. Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble, 2024.

[38] D. M. Ziegler et al. Fine-tuning language models from human preferences. In *Advances in Neural Information Processing Systems*, volume 33, pages 10686–10696, 2020.