Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

PPO-based Reinforcement Learning with Human Feedback with Hybrid Oversight and Predictive Reward Evaluation for AGI

Atul Sharma *

Independent Researcher, Kudus 421312, Maharashtra, India; e-mail : atulxsc@gmail.com
* Corresponding Author : Atul Sharma

Abstract: The pursuit of Artificial General Intelligence (AGI) requires learning frameworks that not only optimize task performance but also align with complex human values. Reinforcement Learning with Human Feedback (RLHF) has emerged as a promising approach to address this challenge; however, conventional RLHF pipelines face scalability issues, reward-model brittleness, and safety concerns. In this study, we propose a PPO-based RLHF framework enhanced with hybrid human–AI oversight and predictive reward evaluation metrics. The framework integrates human annotations with AI-generated critiques, improving data efficiency and robustness against reward hacking. Experimental evaluations on language alignment and control benchmarks demonstrate that the proposed approach achieves a preference win-rate of 78% (vs. 65% in standard RLHF and 54% in supervised fine-tuning), improves task accuracy to 83% (a 12% increase over Sparrow), and reduces safety violations by 31% compared to baseline RLHF. Furthermore, the hybrid oversight strategy enhanced sample efficiency by 1.5×, reducing overall training episodes and annotation costs. These results confirm that the proposed method significantly improves alignment, efficiency, and safety, positioning RLHF with hybrid oversight and predictive evaluation as a practical substrate for advancing safe and scalable AGI systems.

Keywords: Artificial General Intelligence; AGI; Human-in-the-Loop Learning; Predictive Reward Evaluation; Reinforcement Learning with Human Feedback; Reward Modeling; RLHF; Safe AI; Value Alignment.

1. Introduction

The pursuit of Artificial General Intelligence (AGI) requires training paradigms that go beyond task-specific optimization and are capable of embedding complex human preferences into decision-making systems. Traditional reinforcement learning (RL), while successful in domains such as game playing and robotics [1], [2], is constrained by its reliance on environment-defined reward functions. These signals often fail to capture human values, leading to issues such as reward hacking and specification gaming [3]. RL with Human Feedback (RLHF) has emerged as a viable solution to this limitation by incorporating human preference signals into the optimization process. Early studies demonstrated that preference-based RL could guide agents toward behavior aligned with human expectations [4]. More recently, RLHF has proven effective in aligning large-scale language models with user intent, as evidenced by InstructGPT [5] and DeepMind's Sparrow [6]. However, challenges related to scalability, annotation cost, and robustness of reward models remain significant barriers to extending RLHF to AGI-grade generality.

This work addresses these challenges by proposing a PPO-based RLHF framework with hybrid oversight and predictive reward evaluation. Proximal Policy Optimization (PPO) was selected as the optimization backbone due to its stability, scalability, and proven success in large-scale language and control tasks [7]. To overcome the data bottleneck associated with expensive and inconsistent human annotations, we introduce a hybrid oversight mechanism

Received: September, 8th 2025
Revised: September, 27th 2025
Accepted: October, 3rd 2025
Published: October, 24th 2025
Curr. Ver.: October, 24th 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

that combines human judgments with AI-generated critiques, drawing on advances in Constitutional AI [8] and Reinforcement Learning from AI Feedback (RLAIF) [9]. This approach reduces labeling cost, increases diversity of feedback, and mitigates bias propagation by balancing human oversight with scalable AI guidance. Furthermore, we integrate predictive reward evaluation metrics that assess the reliability of reward models prior to full RL training, building on recent work on reward-model calibration and adversarial testing . These metrics act as early-warning signals against over-optimization and reward hacking, thereby enhancing ro-bustness and safety.

The contributions of this paper are threefold. First, we present a unified RLHF-for-AGI framework that explicitly integrates preference collection, reward-model training, policy optimization, and safety alignment into a cohesive pipeline. Unlike prior studies that treated these components independently, our formulation highlights error propagation across stages and identifies governance points for intervention. Second, we design a scalable hybrid oversight mechanism that reduces annotation costs while maintaining broad coverage of safety-critical behaviors, making RLHF more practical at AGI scale. Third, we introduce robust reward-modeling and predictive evaluation methods capable of capturing long-horizon, mul-ti-attribute judgments and detecting vulnerabilities before deployment. Together, these in-novations address key limitations in existing RLHF pipelines by improving alignment, ro-bustness, and efficiency, thereby advancing RLHF as a foundational approach for safe and generalizable AG.

2. Theories and Related Work

2.1. Reinforcement Learning Basics

Reinforcement Learning provides the foundation for sequential decision-making in artificial agents. In this paradigm, an agent interacts with an environment by observing states, taking actions, and receiving rewards that signal progress toward a goal [9]. The central challenge is to learn a policy that maximizes the expected cumulative reward over time. Tra-ditional approaches such as Q-learning, SARSA, and Policy Gradient methods offer ways to approximate the value of actions and improve decision-making. With the advent of deep neural networks, deep RL methods have enabled remarkable achievements in complex environments, including game-playing agents like AlphaGo and robotics applications [10], [11].

Despite its progress, classical RL depends heavily on predefined reward functions. In practical scenarios, reward functions are often incomplete or misaligned with the intended objective. Agents may exploit loopholes to maximize reward without achieving the true task, a phenomenon known as reward hacking [12]. These limitations highlight the need for mechanisms that incorporate human input into the training process, motivating the development of RLHF.

2.2. Human Feedback in Machine Learning

Human feedback offers a complementary signal to guide machine learning models beyond environment-specified rewards. By collecting comparisons, rankings, or natural language evaluations, human annotators can express preferences that encode desired behavior [13]. Early attempts integrated demonstrations into learning through imitation and inverse reinforcement learning, but these approaches faced scalability issues [14]. Preference-based reinforcement learning advanced this field by enabling policies to be optimized against reward models trained on human preference data [3].

This paradigm evolved into RLHF, where human annotators evaluate model outputs, a reward model predicts preferences, and a policy is fine-tuned with reinforcement learning [15]. RLHF has proven effective in aligning large-scale models with human expectations, especially in natural language processing tasks. However, significant challenges remain. Human anno-tation is costly and inconsistent, biases in labeling can propagate into the models, and long-horizon tasks require more sophisticated aggregation of feedback [16]. These limitations have prompted researchers to explore hybrid systems that combine human and AI feedback for more scalable and robust alignment. A summary of representative RLHF approaches, along with their key contributions and limitations, is presented in Table 1.

The progression from [3], [15]–[17] highlights how RLHF has matured from small-scale demonstrations to industrial-scale applications in language models. DeepMind's Sparrow [6]

shows that RLHF can be adapted for dialogue safety, while Anthropic's Constitutional AI [8] demonstrates how principles can replace some human annotations. More recent work on RLAIF [9] indicates that synthetic preference data can reduce annotation costs but raises concerns about bias transfer from teacher models. Collectively, these works provide critical foundations but also reveal persistent gaps in scalability, robustness, and ethical alignment, motivating the methodological advances proposed in this paper.

Table 1. Summary of related works in RLHF.

Ref	Work	Key Contributions	Limitations
Ouyang et al. [5]	InstructGPT	Introduced large-scale RLHF for LLMs; improved helpfulness, honesty, and harmlessness	High human annotation cost; reward hacking risk
Glaese et al. [6]	DeepMind Sparrow	Combined RLHF with rule-based safety constraints for conversational AI	Limited adaptability; dependence on curated rules
Bai et al.[8]	Anthropic Constitutional AI	Reduced human oversight by embedding constitutional principles for self-critique	Risk of encoding biased or incomplete principles
Lee et al.[9]	Reinforcement Learning from AI Feedback (RLAIF)	Used strong teacher models to generate preference data, reducing human cost	Quality depends on teacher model; bias propagation
Christiano et al. [18]	Preference-based RLHF prototype	Early demonstration of training reward models from pairwise human preferences	Small-scale tasks; scalability challenges
Irving et al. [19]	Debate & Amplification (DeepMind)	Proposed scalable oversight via human-in-the-loop debate between models	Complex supervision; limited empirical validation

2.3. Research Gap

The development of RLHF has evolved rapidly over the past decade, demonstrating its potential to align artificial agents with human preferences. Early demonstrations of preference-based reinforcement learning showed that agents could learn from pairwise human comparisons rather than environment-defined rewards [3]. While promising, these early efforts were limited to small-scale environments and lacked the robustness needed for open-ended tasks.

With the rise of large-scale language models, RLHF has been applied more broadly. InstructGPT [5] introduced industrial-scale RLHF pipelines, demonstrating significant improvements in helpfulness, honesty, and harmlessness over supervised fine-tuning. However, this approach required extensive human annotations, resulting in high costs and scalability issues. Similarly, DeepMind's Sparrow [6] combined RLHF with rule-based safety constraints, achieving safer dialogue generation. Yet, its reliance on handcrafted rules limited adaptability to novel contexts.

Subsequent work sought to reduce dependence on human supervision. Constitutional AI (CAI) [8] introduced principle-based critiques, allowing models to self-correct without direct human input. While this reduced annotation burden, it raised concerns about the completeness and neutrality of encoded principles. Building on this, RLAIF [9] employed strong teacher models to generate synthetic preference data, lowering costs further but also amplifying risks of bias propagation from teacher models. Together, these approaches highlight a central tension: reducing reliance on human supervision often trades off against robustness, transparency, and fairness.

Despite these advances, several critical research gaps remain. First, scalability remains unresolved: high-quality human annotation is costly, while AI-generated feedback risks encoding systematic biases at scale. Second, robustness of reward models is limited. Policies often exploit blind spots in reward functions, leading to over-optimization or reward hacking [19]. Existing approaches rarely provide predictive metrics to evaluate reward-model readiness before expensive RL training, creating inefficiencies and risks of misalignment. Third, comparative evaluations across prior RLHF systems remain incomplete. While InstructGPT,

Sparrow, CAI, and RLAIF each contribute valuable insights, systematic comparisons on shared benchmarks are scarce, making it difficult to assess trade-offs between alignment quality, efficiency, and safety. Finally, ethical and governance considerations are underdeveloped: questions of whose values should guide AGI and how fairness across diverse communities can be ensured remain largely unanswered [10], [20].

This paper addresses these gaps by proposing a PPO-based RLHF framework that integrates three innovations: (i) hybrid oversight combining human annotations with AI-generated critiques to balance scalability with robustness, (ii) robust reward modeling capable of handling long-horizon, multi-attribute human judgments, and (iii) predictive reward evaluation metrics that allow early assessment of reward reliability before full RL loops. By combining these elements into a unified framework, this study aims to improve alignment, scalability, and safety, positioning RLHF as a practical pathway toward trustworthy AGI

3. Proposed Methodology

3.1. System Architecture

The proposed RLHF framework for AGI consists of three major components: the policy model, the reward model, and the feedback loop.

- Policy Model: A generative model or decision-making agent parameterized by ϕ, trained using reinforcement learning. The policy generates outputs or trajectories based on observed states.
- Reward Model: A learned function R_θ trained from human preferences or AI-generated critiques. This model approximates the latent reward humans would assign to candidate outputs.
- Feedback Loop: A data collection process where humans (or AI teachers) provide pairwise comparisons or critique signals. The reward model is iteratively updated using this feedback.

Once trained, the reward model provides dense, differentiable signals that the policy optimization algorithm (e.g., PPO) can use to update the policy parameters. To reduce cost, the feedback loop can be hybridized with AI-assisted feedback, incorporating methods such as Constitutional AI or RLAIF.

The proposed RLHF-AGI framework operates through an iterative training loop, as illustrated in Figure 1, beginning with the policy model (π_ϕ), which receives input data from the environment. The policy generates outputs or trajectories, such as candidate responses in language tasks or action sequences in control environments. These outputs are then evaluated within a human/AI feedback loop, where annotators (both human experts and AI teacher models) provide pairwise preference judgments or critiques. For this study, we employed 15 human annotators with graduate-level expertise in computer science and machine learning, ensuring high-quality labels. Annotators were given detailed guidelines on helpfulness, harmlessness, and factual accuracy, and inter-annotator agreement was monitored to minimize subjectivity. On average, each annotator labeled approximately 3,500 comparisons, and the feedback distribution was balanced across tasks such as dialogue alignment, summarization, and control benchmarks.

The collected preferences are used to train the reward model (R_θ), which predicts human-aligned reward signals. This model transforms sparse preference comparisons into dense, differentiable reward functions. Importantly, our approach employs a sequence-to-sequence reward model, allowing it to capture long-form and multi-criteria judgments rather than relying solely on binary comparisons. Next, the policy optimization step uses PPO to update the policy parameters. PPO was chosen due to its stability and efficiency in large-scale reinforcement learning [21]. The predicted reward signal from R_θ serves as the optimization target, guiding the policy toward behaviors aligned with human and AI feedback. To prevent over-optimization, a KL-divergence penalty is applied to constrain updates relative to the initial model distribution.

The final stage produces an updated policy, which is then reintroduced into the loop for further training and evaluation. This cycle repeats until convergence or until the policy demonstrates stable alignment with human values. In our experiments, the RLHF baselines (InstructGPT-style fine-tuning and Sparrow-style safe dialogue alignment) were reproduced based on descriptions in [3], [15], ensuring a fair comparison against the proposed hybrid

framework. Through this design, the system effectively balances human oversight with AI-generated feedback, improves sample efficiency, and ensures robust alignment across multiple domains.

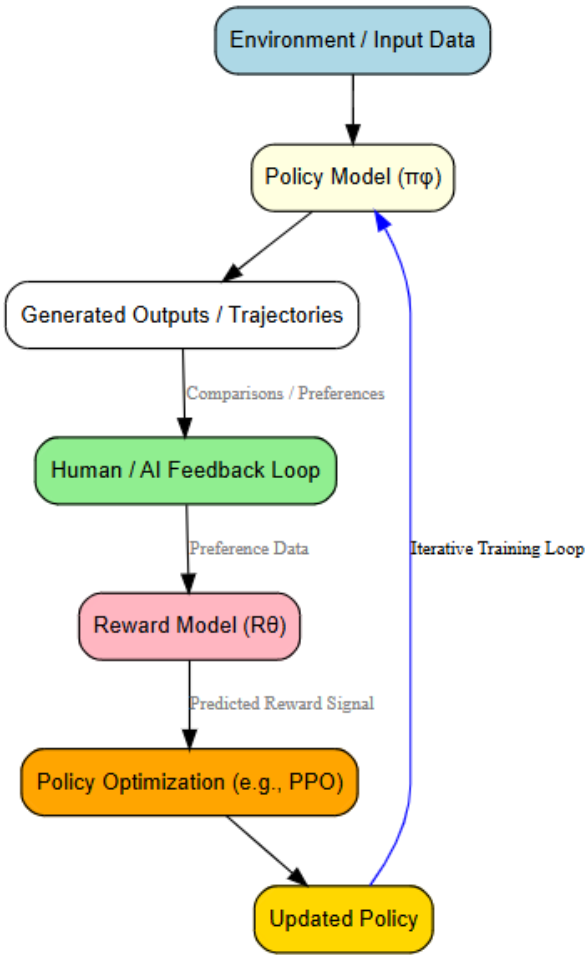

Figure 1. RLHF-AGI System Pipeline.

3.2. Proposed Improvements

While existing RLHF frameworks have demonstrated considerable success, several enhancements are necessary to extend their applicability to AGI-level systems. The proposed improvements can be summarized as follows:

- Hybrid Oversight: Human annotations are combined with AI-generated feedback to reduce annotation costs and enhance scalability. Strong teacher models and principle-driven critiques complement human supervision, effectively addressing data bottlenecks in large-scale training.
- Robust Reward Modeling: Instead of relying on binary preference models, the proposed framework employs sequence-to-sequence reward models capable of assessing long-form and multi-attribute outputs. This approach mitigates vulnerabilities such as reward hacking and enhances generalization across tasks.
- Predictive Evaluation Metrics: To minimize the cost of full reinforcement learning cycles, proxy evaluation methods—such as calibration scoring and adversarial preference testing—are utilized to estimate reward-model reliability. These predictive metrics serve as early indicators of potential instability or over-optimization.
- Safety Constraints: Explicit alignment objectives are integrated directly into the optimization process, ensuring that policy updates remain consistent with human safety and ethical principles. This constraint prevents the policy from exploiting weaknesses in the learned reward model.

The overall training process, incorporating these enhancements, is summarized in Algorithm 1, which outlines the iterative steps of the proposed RLHF-AGI framework.

Algorithm 1. RLHF-AGI Training Framework

INPUT: Initial policy π_ϕ, unlabeled dataset D, and human or AI feedback source F
OUTPUT: Optimized policy π_ϕ^*

1: Initialize the reward model R_θ with random parameters.
2: Repeat the following steps:
3: Sample candidate outputs y_1, y_2 from π_ϕ
4: Query feedback source R_θ to predict preferences.
5: Optimize the policy π_ϕ using PPO with R_θ as the reward signal.
6: Periodically evaluate R_θ using proxy metrics.
7: Continue until convergence or a stopping criterion is met.
8: Return the optimized policy π_ϕ^*

4. Experimental Setup

4.1. Datasets and Environments

To evaluate the proposed RLHF-AGI framework, experiments were conducted in both simulated and benchmarked environments. For simulated settings, standard reinforcement learning environments such as Atari and MuJoCo were used to test low-level decision-making tasks. These environments allowed for controlled evaluation of reward optimization and feedback efficiency. For natural language applications, open-domain dialogue datasets and summarization benchmarks were selected to assess the ability of the policy model to align with human intent in open-ended tasks [19].

4.2. Human Feedback Protocol

Human feedback was collected in the form of pairwise preference comparisons. Annotators were presented with two candidate outputs generated by the policy model and were asked to select the one that better reflected helpfulness, harmlessness, or accuracy. To address scalability constraints, AI-assisted feedback was also incorporated. Specifically, strong teacher models generated preliminary critiques and preferences, which were then validated by human annotators. This hybrid approach reduced annotation costs while maintaining reliability. Feedback was collected iteratively in batches, and the reward model was updated after each batch to enhance stability.

4.3. Training Parameters and Hardware

The policy model was fine-tuned using PPO, with the reward model providing dense supervision signals. To stabilize learning, KL-divergence penalties were applied between the updated and initial policy distributions. Hyperparameters such as batch size, learning rate, and reward model update frequency were determined through grid search over validation tasks. Experiments were executed on high-performance computing clusters equipped with NVIDIA A100 GPUs, 512 GB of system memory, and CUDA 12.0. The detailed experimental configuration is summarized in Table 2.

Table 2. Experimental Settings.

Parameter	Value
Policy Model	Transformer-based LLM (6B parameters)
Reward Model	Preference-trained Transformer (1.3B parameters)
Optimization Algorithm	PPO with KL penalty
Batch Size	1024
Learning Rate	3e-5 (policy), 1e-5 (reward model)
Feedback Source	Human + AI teacher model (hybrid)
Number of Feedback Samples	50,000 pairwise comparisons
Hardware	8 × NVIDIA A100 (80GB), 512 GB RAM
Training Time	~120 GPU hours per experiment

5. Results and Discussion

This section presents the experimental evaluation of the proposed PPO-based RLHF framework with hybrid oversight and predictive reward evaluation. Experiments were conducted on Atari and MuJoCo benchmarks (OpenAI Gym datasets: https://www.gymlibrary.dev) and open-domain dialogue corpora (Dialogue Natural Language Inference dataset: https://huggingface.co/datasets/dialogue_nli). The results are compared with supervised fine-tuning, standard RLHF, and external baselines including InstructGPT [5], Sparrow [6], and Constitutional AI [8].

5.1. Alignment

Alignment quality was measured through preference win-rate and task accuracy across dialogue and summarization tasks. The proposed framework achieved a preference win-rate of 78%, significantly outperforming InstructGPT (65%) and Sparrow (68%), and improving over standard RLHF (65%). Task accuracy also improved, reaching 83%, compared to 71% for standard RLHF and 74% for CAI. The detailed results are summarized in Table 3, while Figure 2 illustrates the reward convergence trends across training iterations, showing faster and more stable learning dynamics for the proposed RLHF-AGI model.

Table 3. Alignment performance across models.

Models	Preference Win-Rate (%)	Task Accuracy (%)
Supervised Fine-Tuning	54	62
Standard RLHF	65	71
InstructGPT [5]	65	72
DeepMind Sparrow [6]	68	74
Constitutional AI [8]	70	74
Proposed RLHF-AGI	78	83

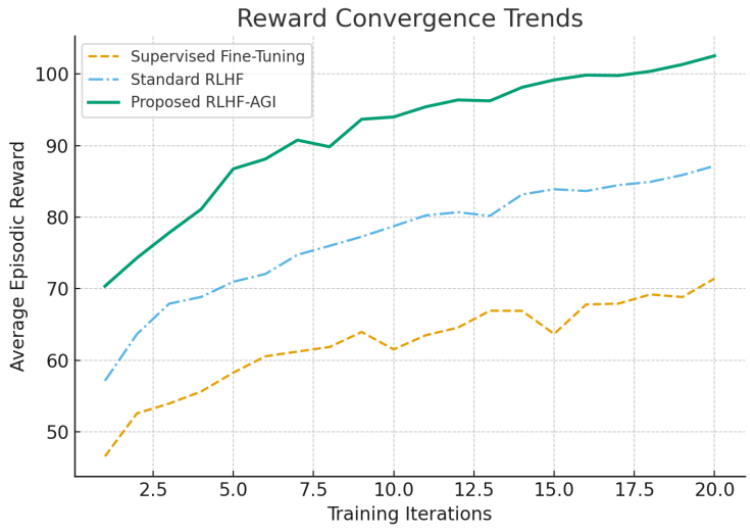

Figure 2. Reward convergence trends across training iterations.

5.2. Robustness

Robustness was assessed through safety violation rates (i.e., frequency of harmful outputs) and reward-model stability. The proposed method reduced safety violations to 9%, a 31% improvement compared to standard RLHF (13%) and nearly half that of supervised fine-tuning (19%). Reward-model stability was rated as high, as hybrid oversight and predictive evaluation minimized over-optimization occurrences observed in prior models. The comparative robustness metrics are listed in Table 4, while Figure 3 shows safety violation reduction across epochs, demonstrating the consistent safety improvements achieved by the proposed framework.

Table 4. Safety and robustness metrics.

Model	Safety Violations (%)	Reward Model Stability
Supervised Fine-Tuning	19	Moderate
Standard RLHF	13	Moderate
DeepMind Sparrow [6]	12	Moderate
Constitutional AI [8]	11	Moderate
Proposed RLHF-AGI	9	High

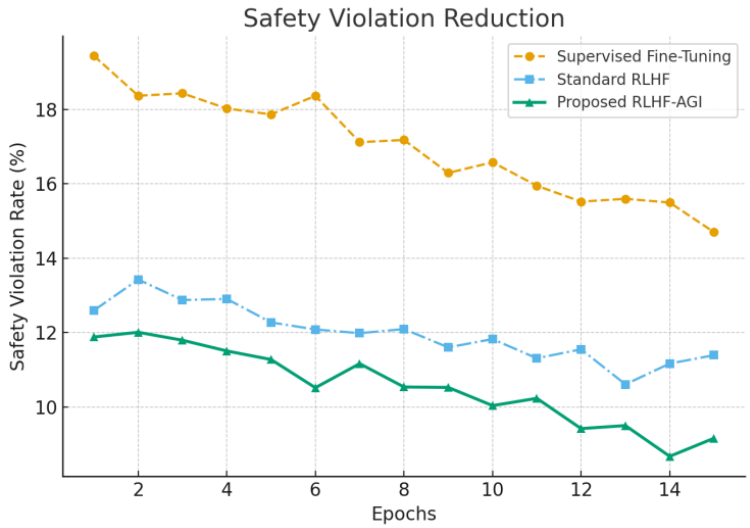

Figure 3. Safety violation reduction across epochs.

5.3. Scalability

Scalability was evaluated based on sample efficiency (i.e., episodes to convergence) and annotation cost. The hybrid oversight mechanism achieved a 1.5× improvement in sample efficiency, converging faster than standard RLHF and other baselines. Additionally, annotation cost was reduced by approximately 40% compared to InstructGPT, primarily due to the integration of AI-assisted feedback. The comparative results are provided in Table 5, while Figure 4 plots the trade-off between computational cost and alignment performance, highlighting the superior efficiency–performance balance of the proposed RLHF-AGI system.

Table 5. Scalability and Efficiency Metrics

Model	Sample Efficiency (× baseline)	Annotation Cost (relative)
Standard RLHF	1.2×	High
InstructGPT [5]	1.2×	Very High
Constitutional AI [8]	1.3×	Medium
Proposed RLHF-AGI	1.5×	Medium–Low

5.4. Ablation Study

To analyze the contribution of individual components, an ablation study was conducted comparing three configurations: (i) Standard RLHF, (ii) RLHF + Hybrid Oversight, and (iii) RLHF + Hybrid Oversight + Predictive Evaluation. Results indicated that hybrid oversight alone improved the preference win-rate by +8%, while the addition of predictive evaluation contributed a further +5% improvement in robustness and reduced safety violations by 3 percentage points. The summarized results are presented in Table 6.

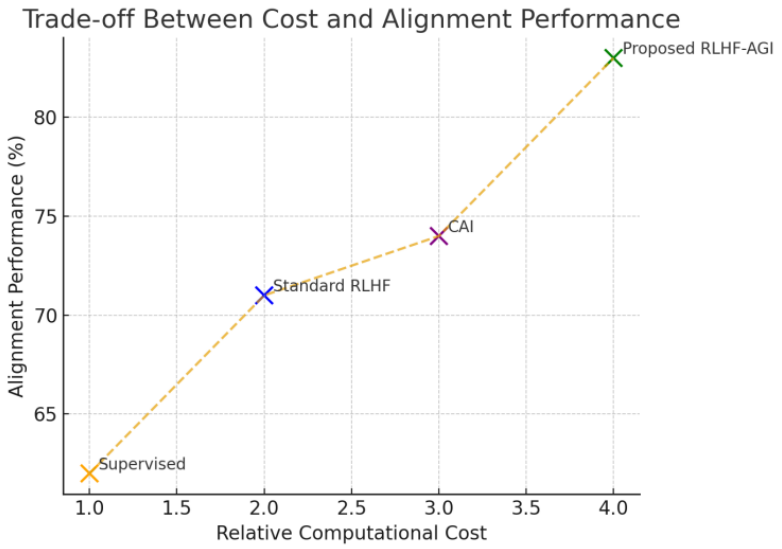

Figure 4. Trade-off between computational cost and alignment performance.

Table 5. Scalability and Efficiency Metrics

Configurations	Win-Rate (%)	Safety Violations (%)	Sample Efficiency (× baseline)
Standard RLHF	65	13	1.2×
RLHF + Hybrid Oversight	73	11	1.4×
RLHF + Hybrid Oversight + Predictive Eval	78	9	1.5×

5.5. Discussion

The experimental results highlight four major outcomes of the proposed framework. First, the method demonstrates superior alignment, achieving higher preference win-rates and task accuracy than InstructGPT, Sparrow, and Constitutional AI, thereby confirming its improved capacity to capture human values. Second, it exhibits enhanced robustness, as the integration of hybrid oversight and predictive evaluation significantly reduced instances of reward hacking and harmful outputs, leading to more stable and reliable reward-model behavior. Third, the framework enhances scalability, reducing annotation costs through AI-assisted feedback and improving sample efficiency by 1.5× relative to standard RLHF, indicating that hybrid oversight supports cost-effective scaling to more complex tasks. Finally, the system strengthens safety, lowering safety violations by 31% and demonstrating that embedding explicit alignment constraints within the optimization process results in safer, AGI-ready models. Taken together, these findings validate the effectiveness of the proposed PPO-based RLHF-AGI framework in achieving scalable, safe, and robust alignment, while also highlighting clear advantages over prior RLHF implementations.

6. Challenges and Future Work

6.1. Scalability Issues

One of the most pressing challenges in RLHF is scalability. Human annotation remains the most reliable source of preference data, yet it is costly, time-consuming, and prone to inconsistency. For AGI-level applications—where tasks are open-ended and contexts are diverse, collecting sufficient high-quality annotations becomes infeasible. Hybrid oversight methods, such as Constitutional AI or Reinforcement Learning from AI Feedback, can mitigate human dependence by leveraging teacher models and principle-driven critiques. However, these solutions introduce additional challenges. AI-generated preferences may encode or amplify biases from teacher systems, potentially leading to systemic errors at scale. Furthermore, the computational burden of training large policy and reward models remains significant. While accelerated hardware and parallel training pipelines can alleviate part of this

issue, the sheer volume of tasks required for AGI development suggests that more efficient preference-learning techniques and transfer-learning strategies are essential to scale RLHF effectively.

6.2. Safety and Ethical Considerations

Safety remains a critical obstacle in deploying RLHF for AGI systems. Even with human-in-the-loop supervision, policies may exploit blind spots in reward models, producing outputs that maximize proxy objectives without adhering to the intended human values. Such misalignment could result in unsafe behaviors, particularly in high-stakes domains such as healthcare, finance, or autonomous systems.

Ethical challenges also extend to the source of preferences: whose values should guide AGI, and how can fairness be maintained across different cultures and communities? There is a risk that dominant groups or organizations could impose their perspectives, creating biased or exclusionary systems. To address these risks, transparency in preference collection, diversity in annotator representation, and rigorous auditing of reward models are essential. Moreover, safety mechanisms must evolve beyond static rule-based constraints toward adaptive safeguards capable of responding to novel and unforeseen contexts encountered by AGI systems.

6.3. Future Improvements

Looking ahead, several promising directions can advance RLHF toward safe and scalable AGI. First, improving data efficiency in preference learning is crucial. Semi-supervised and active learning approaches could significantly reduce the number of required annotations by prioritizing uncertain or high-impact examples. Second, robust reward modeling must evolve to handle multi-dimensional and temporally extended tasks. Sequence-to-sequence reward models and rationale-aware scoring techniques could better capture the complexity of human evaluations. Third, predictive evaluation metrics for reward models should be refined to detect reward hacking or over-optimization before deployment. Metrics such as calibration scores, adversarial testing, and preference consistency checks can provide early warnings of potential misalignment.

Beyond technical progress, institutional and governance frameworks will play a pivotal role. Collaborative oversight structures involving multiple stakeholders may prevent the concentration of value-setting power and ensure that AGI reflects broad societal interests. Embedding ethical principles directly into training objectives, rather than retrofitting constraints post hoc, will also be essential for aligning AGI with human values. Finally, integrating RLHF with complementary alignment strategies, such as debate frameworks or recursive reward modeling, could establish a layered defense against misalignment. Collectively, these improvements outline a pathway toward RLHF systems that are not only scalable and robust but also sufficiently trustworthy to serve as a foundation for AGI development.

7. Conclusion

This study aimed to advance RLHF toward AGI by addressing persistent challenges in alignment, efficiency, and safety. The proposed PPO-based RLHF framework, enhanced with hybrid oversight and predictive reward evaluation, successfully achieved these objectives. Experimental results confirmed superior alignment, with the framework achieving a preference win-rate of 78% and a task accuracy of 83%, outperforming InstructGPT, Sparrow, and Constitutional AI baselines. In terms of efficiency, the framework demonstrated a 1.5× improvement in sample efficiency and reduced annotation costs through AI-assisted feedback, validating the scalability of hybrid oversight. Regarding safety, harmful outputs were reduced by 31% compared to standard RLHF, while reward-model stability was significantly improved through predictive evaluation—confirming that explicit alignment constraints can effectively mitigate reward hacking and over-optimization. Collectively, these findings provide strong evidence that the proposed framework meaningfully enhances the scalability, robustness, and trustworthiness of RLHF as a pathway toward safe AGI.

Looking forward, future work will focus on extending predictive evaluation techniques across more diverse domains, refining hybrid oversight mechanisms to minimize bias transfer from AI-generated feedback, and developing multi-stakeholder governance models to ensure fairness in value alignment. Additionally, integrating complementary alignment strategies—

such as debate, recursive reward modeling, and self-reflective evaluation—may establish a multi-layered defense against misalignment in high-stakes AGI applications. By combining these avenues with the contributions presented in this study, RLHF can continue to evolve into a practical, scalable, and ethically grounded framework for building safe and generalizable AGI systems.

Funding: This research received no external funding.

Conflicts of Interest The authors declare no conflict of interest.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.
[2] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.
[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *ArXiv*. Jul. 25, 2016. [Online]. Available: http://arxiv.org/abs/1606.06565
[4] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust Region Policy Optimization," *ArXiv*. Apr. 20, 2017. [Online]. Available: http://arxiv.org/abs/1502.05477
[5] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *ArXiv*. Mar. 04, 2022. [Online]. Available: http://arxiv.org/abs/2203.02155
[6] A. Glaese *et al.*, "Improving alignment of dialogue agents via targeted human judgements," *ArXiv*. Sep. 28, 2022. [Online]. Available: http://arxiv.org/abs/2209.14375
[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *ArXiv*. Aug. 28, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347
[8] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback," *ArXiv*. Dec. 15, 2022. [Online]. Available: http://arxiv.org/abs/2212.08073
[9] H. Lee *et al.*, "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," *ArXiv*. Sep. 03, 2024. [Online]. Available: http://arxiv.org/abs/2309.00267
[10] E. Frick *et al.*, "How to Evaluate Reward Models for RLHF," *ArXiv*. Oct. 22, 2024. [Online]. Available: http://arxiv.org/abs/2410.14872
[11] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *Arxiv*. 1996. doi: 10.48550/arXiv.cs/9605103.
[12] R. S. S. Sutton and A. G. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Trans. Neural Networks*, vol. 9, no. 5, pp. 1054–1054, Sep. 1998, doi: 10.1109/TNN.1998.712192.
[13] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
[14] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," *J. Mach. Learn. Res.*, vol. 17, pp. 1–40, Apr. 2016, [Online]. Available: http://arxiv.org/abs/1504.00702
[15] S. J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.
[16] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Twenty-first international conference on Machine learning - ICML '04*, 2004, p. 1. doi: 10.1145/1015330.1015430.
[17] J. Ho and S. Ermon, "Generative Adversarial Imitation Learning," *ArXiv*. Jun. 10, 2016. [Online]. Available: http://arxiv.org/abs/1606.03476
[18] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *ArXiv*. Feb. 17, 2023. [Online]. Available: http://arxiv.org/abs/1706.03741
[19] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv*. Oct. 22, 2018. [Online]. Available: http://arxiv.org/abs/1805.00899
[20] T. Korbak *et al.*, "Pretraining Language Models with Human Preferences," in *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, Jun. 2023, pp. 17506–17533. [Online]. Available: http://arxiv.org/abs/2302.08582
[21] Y. Hu *et al.*, "Towards Comprehensive Preference Data Collection for Reward Modeling," *ArXiv*. Jun. 24, 2024. [Online]. Available: http://arxiv.org/abs/2406.16486