A Survey of Reinforcement Learning Techniques and Applications

www.surveyx.cn

Abstract

Reinforcement Learning (RL) is a pivotal machine learning paradigm that trains agents to make sequential decisions by maximizing cumulative rewards. This survey paper explores the core concepts and recent advancements in RL, focusing on reward modeling, policy gradient methods, and the integration of human feedback. Key topics include Proximal Policy Optimization (PPO) and Generalized Proximal Policy Optimization (GRPO), which enhance training stability and efficiency. The survey also examines the role of large-scale language models (LLMs) in RL, highlighting their potential to process complex linguistic inputs and improve decision-making processes. The integration of human feedback through Reinforcement Learning from Human Feedback (RLHF) is emphasized as crucial for aligning AI systems with human values. Challenges in designing reward functions, enhancing sample efficiency, and balancing exploration and exploitation are discussed, alongside strategies for optimizing computational resources. The paper concludes by outlining future directions in RL research, advocating for the development of more robust, adaptable, and human-aligned AI systems. By addressing existing challenges and leveraging novel methodologies, RL can continue to drive significant advancements in artificial intelligence, leading to more sophisticated and human-centric solutions.

1 Introduction

1.1 Overview of Reinforcement Learning

Reinforcement Learning (RL) is a dynamic machine learning framework where agents learn to make sequential decisions aimed at maximizing cumulative rewards through continuous interaction with their environments. This iterative process involves agents perceiving their state, executing actions, and receiving feedback in the form of rewards, which informs future decisions. The feedback loop inherent to RL distinguishes it from other paradigms, enabling the resolution of complex decision-making challenges that traditional programming struggles to address, particularly in high-dimensional state and action spaces [1].

The significance of RL spans various domains, each presenting distinct challenges and opportunities. In healthcare, RL supports the development of systems that prioritize patient safety and optimize treatment outcomes through risk-aware decision-making frameworks [2]. In autonomous driving, RL algorithms are crucial for navigating the unpredictability and dynamics of real-world environments, thereby advancing intelligent driving systems [3]. Moreover, the integration of RL with Large Language Models (LLMs) has markedly improved the alignment of AI outputs with human preferences, enhancing the quality of human-computer interactions [4].

A pivotal aspect of RL is the incorporation of human feedback, which enhances adaptability and efficiency in online learning approaches, addressing the limitations of traditional offline methods. This integration allows RL systems to better align with human values and preferences [5]. Additionally,

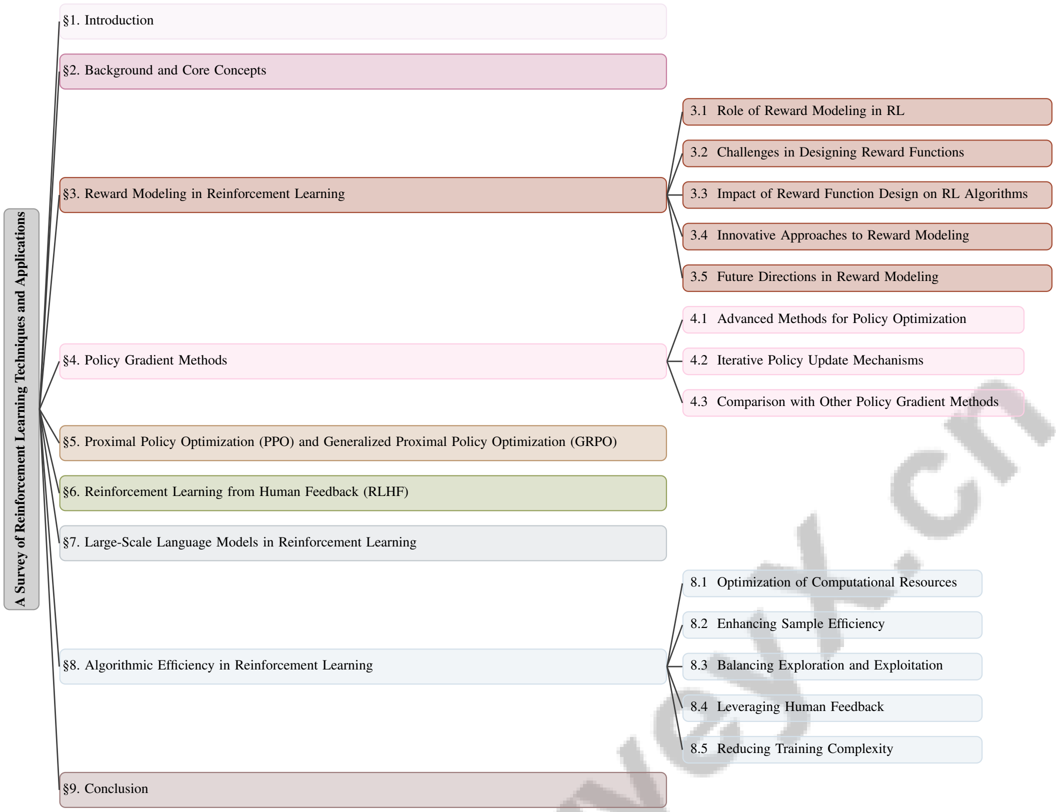

A Survey of Reinforcement Learning Techniques and Applications
§1. Introduction
§2. Background and Core Concepts
§3. Reward Modeling in Reinforcement Learning
3.1 Role of Reward Modeling in RL
3.2 Challenges in Designing Reward Functions
3.3 Impact of Reward Function Design on RL Algorithms
3.4 Innovative Approaches to Reward Modeling
3.5 Future Directions in Reward Modeling
§4. Policy Gradient Methods
4.1 Advanced Methods for Policy Optimization
4.2 Iterative Policy Update Mechanisms
4.3 Comparison with Other Policy Gradient Methods
§5. Proximal Policy Optimization (PPO) and Generalized Proximal Policy Optimization (GRPO)
§6. Reinforcement Learning from Human Feedback (RLHF)
§7. Large-Scale Language Models in Reinforcement Learning
§8. Algorithmic Efficiency in Reinforcement Learning
8.1 Optimization of Computational Resources
8.2 Enhancing Sample Efficiency
8.3 Balancing Exploration and Exploitation
8.4 Leveraging Human Feedback
8.5 Reducing Training Complexity
§9. Conclusion

Figure 1: chapter structure

RL's ability to optimize non-differentiable objectives in deep neural networks has been instrumental for tasks requiring sophisticated model tuning and performance optimization [6].

The versatility of RL is further illustrated in competitive tasks, where hierarchical structures with multiple cooperative critics can enhance learning processes and increase cumulative rewards [7]. In physics-based domains, RL has proven effective in evolving algorithms and applications, particularly in 2D object manipulation tasks [8]. However, the inefficiency of current RL methods, which require extensive computational resources and exhibit sample inefficiency, presents challenges in learning effective control policies due to their slow and unstable nature [9].

The extensive application of RL across diverse sectors, particularly in enhancing LLMs through techniques such as Reinforcement Learning from Human Feedback (RLHF) and advanced credit assignment methods like VinePPO, underscores its transformative potential in artificial intelligence research and development. This potential is highlighted by RL's ability to improve model performance, mitigate issues such as toxicity and hallucinations, and refine reasoning capabilities in complex tasks, despite the inherent challenges associated with its implementation and the necessity for a comprehensive understanding of its methodologies [10, 11, 12].

1.2 Structure of the Survey

This survey provides a comprehensive examination of key aspects and recent advancements in Reinforcement Learning (RL). It begins with an introduction to RL as a unique machine learning paradigm, emphasizing its feedback loop mechanism and applications across various domains. The foundational concepts and approaches within RL are thoroughly examined, essential for understanding advanced topics such as Reinforcement Learning from Human Feedback (RLHF) and their implications for Large Language Models (LLMs). This section clarifies the complexities of RL techniques, including reward modeling and optimization strategies, establishing a solid groundwork for researchers navigating the challenges and advancements in RL-enhanced LLMs [10, 13, 12].

AI-generated, for reference only.

The second section explores core concepts of RL, including Reward Modeling, Proximal Policy Optimization (PPO), Generalized Proximal Policy Optimization (GRPO), and RLHF. It also covers policy gradient methods, large-scale language models, and algorithmic efficiency, providing a robust theoretical framework for subsequent discussions.

The third section delves into Reward Modeling, discussing its critical role in guiding decision-making processes within RL. It examines the challenges of designing effective reward functions and their impact on RL algorithm performance, while highlighting innovative approaches and potential future directions.

In the fourth section, policy gradient methods are investigated, focusing on their capability to optimize policies by estimating expected reward gradients. Advanced methods for policy optimization, iterative policy updates, and comparisons with other optimization techniques are discussed.

The fifth section evaluates PPO and GRPO, advanced policy gradient methods that enhance training stability and efficiency in RL. Applications, use cases, challenges, and potential future directions are explored.

The sixth section analyzes RLHF, emphasizing its importance in aligning AI models with human values and the challenges of integrating human feedback. Scalability, efficiency, and the complexities of human feedback integration in RL environments are discussed.

The seventh section examines the application of large-scale language models in RL, focusing on their role in understanding and generating human-like language. Integration techniques, prompt optimization, evaluation, and benchmarking within RL applications are explored.

The eighth section highlights algorithmic efficiency in RL, discussing strategies for optimizing computational resources, enhancing sample efficiency, balancing exploration and exploitation, leveraging human feedback, and reducing training complexity.

The conclusion synthesizes primary insights from the literature on RL in the context of LLMs, emphasizing current advancements and challenges. It addresses the complexities of implementing RL techniques, such as RLHF and PPO, which are critical for enhancing LLM capabilities in complex reasoning tasks. The analysis also considers limitations of existing approaches, particularly in credit assignment and reward modeling, while suggesting potential future research directions to further advance RL integration in LLM development [10, 13, 14, 15, 12].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Foundational Concepts and Variants

Reinforcement Learning (RL) is a critical machine learning field focused on training agents to maximize cumulative rewards through sequential decision-making. Central to RL is the Markov Decision Process (MDP), a mathematical framework for decision-making in stochastic environments, which is crucial for assessing RL algorithm effectiveness, especially in deterministic settings [6].

RL algorithms are categorized into value-based, policy-based, model-based, and model-free methods. Value-based approaches, like Q-learning, estimate action values to indirectly derive optimal policies, while policy-based methods such as Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO) directly optimize policies. The implementation nuances in deep RL, particularly for PPO and TRPO, are vital for robust performance in environments with delayed rewards [16, 11].

Model-based methods leverage environmental models for predicting future states and rewards, offering sample-efficient learning. In contrast, model-free methods, though simpler, grapple with the exploration-exploitation dilemma—balancing new state exploration against exploiting known rewards to avoid suboptimal policies [17].

Recent advances include preference-based reinforcement learning, which uses preferences between state-action pairs instead of absolute rewards, advantageous in scenarios with complex reward quantification [18]. The integration of RL with large language models (LLMs) exposes limitations in traditional methods like next-token prediction, necessitating innovative strategies for leveraging LLM capabilities in RL [4].

A significant challenge in RL is effectively learning multimodal policies, which identify multiple optimal solutions in complex environments. Current deep reinforcement learning (DRL) algorithms often fall short, leading to suboptimal learning efficiency. Additionally, quantifying uncertainty in on-policy actor-critic methods is crucial for detecting out-of-distribution states during inference [19].

Exploring risk-aware objectives in RL is gaining traction, focusing on one-episode-reward settings where traditional algorithms may not directly apply [1]. Addressing these challenges is essential for developing robust RL solutions, with foundational concepts providing a framework for advancing sophisticated algorithms and applications.

2.2 Innovative Approaches and Algorithms

The evolution of Reinforcement Learning (RL) is characterized by innovative approaches and algorithms aimed at enhancing RL systems' efficiency and effectiveness. Pairwise Proximal Policy Optimization (P3O) represents a significant advancement, optimizing large language models (LLMs) using relative feedback from comparative rewards, thereby addressing limitations of traditional policy gradient methods [19].

To reduce high variance and empirical bias in chaotic environments, Gradient-Informed Proximal Policy Optimization (GIPPO) refines the RP gradient, enhancing the stability and performance of policy gradient methods [20]. Behavior Proximal Policy Optimization (BPPO) improves behavior policy by leveraging PPO's intrinsic properties without additional constraints, streamlining optimization [21].

Incorporating predictive processing mechanisms into RL algorithms, as demonstrated by Predictive Processing Proximal Policy Optimization (P4O), significantly enhances learning efficiency by anticipating future states and rewards, particularly beneficial in resource-constrained environments due to improved sample efficiency [9].

In Safe Reinforcement Learning (SRL), the Adaptive Primal-Dual (APD) method dynamically adjusts learning rates based on Lagrangian multipliers to optimize policy while ensuring compliance with safety constraints [22].

Preference-based reinforcement learning has led to joint demonstration and preference learning methods, enhancing diagnostic accuracy by merging traditional RL techniques with modern data-driven insights [18]. These innovations demonstrate the potential of integrating RL with contemporary machine learning advancements to address complex decision-making challenges.

Further developments, such as reinforcement learning from human feedback (RLHF) and online iterative RLHF, reflect RL research's dynamic nature, targeting improvements in LLMs' performance and alignment with human preferences while addressing reward modeling and algorithmic implementation challenges. Tools like RLInspect provide deeper insights into RL model behavior, aiding researchers in navigating training and evaluation complexities [10, 23, 24, 12]. These advancements collectively underscore the ongoing commitment to enhancing RL capabilities and resolving existing limitations across diverse domains.

In recent years, the field of reinforcement learning (RL) has witnessed significant advancements, particularly in the area of reward modeling. A comprehensive understanding of this domain necessitates an exploration of its hierarchical structure, which encompasses various roles, challenges, innovative approaches, and future directions. As illustrated in Figure 2, this figure categorizes the main concepts into five primary sections, each accompanied by subcategories and detailed points. This organization emphasizes the critical importance of reward function design and its profound impact on the efficacy of RL algorithms. By examining these elements, we can gain insights into the complexities and nuances that define reward modeling in RL.

3 Reward Modeling in Reinforcement Learning

3.1 Role of Reward Modeling in RL

Reward modeling is pivotal in Reinforcement Learning (RL), enabling agents to assess the value of their actions within an environment. Reinforcement Learning from Human Feedback (RLHF) transforms human preference judgments into actionable learning signals, refined by fine-grained feedback to enhance model performance and alignment with human expectations [25, 26, 27, 28].

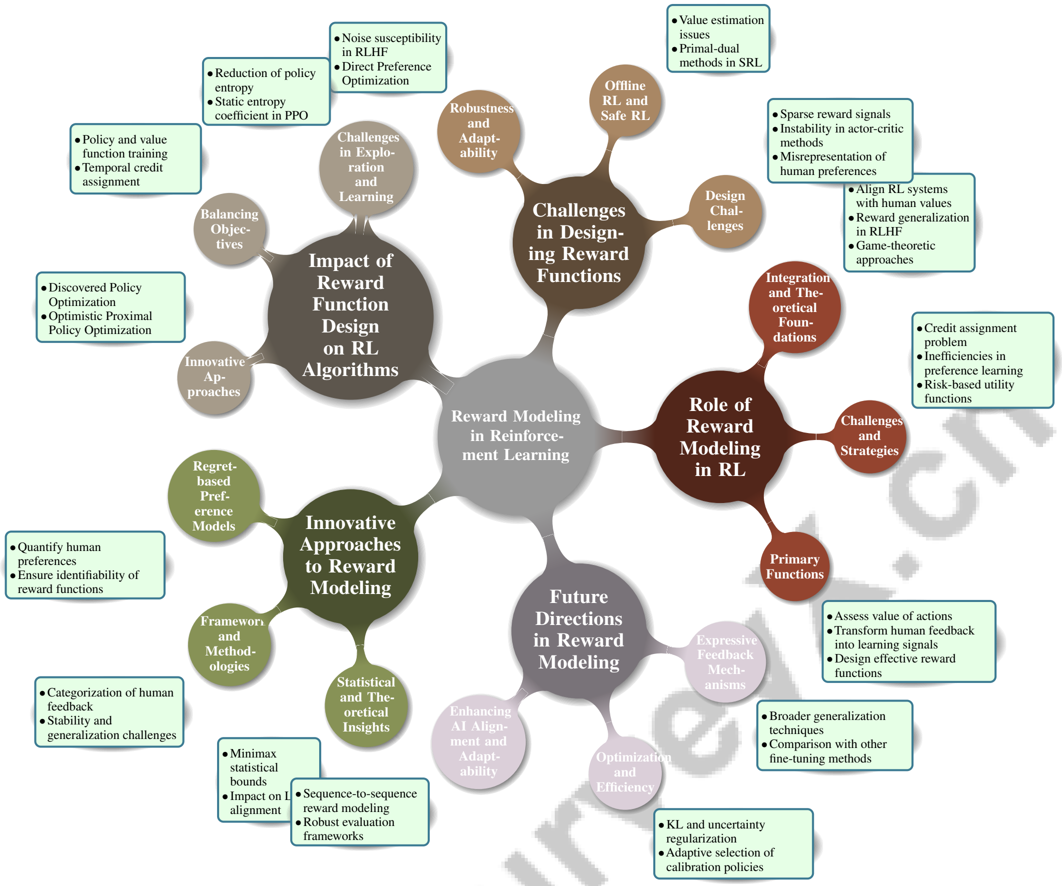

Reward Modeling in Reinforcement Learning
Impact of Reward Function Design on RL Algorithms
Challenges in Exploration and Learning
Balancing Objectives
Innovative Approaches
• Noise susceptibility in RLHF
• Direct Preference Optimization
• Reduction of policy entropy
• Static entropy coefficient in PPO
• Policy and value function training
• Temporal credit assignment
• Discovered Policy Optimization
• Optimistic Proximal Policy Optimization
Challenges in Designing Reward Functions
Robustness and Adaptability
Offline RL and Safe RL
Design Challenges
• Value estimation issues
• Primal-dual methods in SRL
• Sparse reward signals
• Instability in actor-critic methods
• Misrepresentation of human preferences
• Align RL systems with human values
• Reward generalization in RLHF
• Game-theoretic approaches
Role of Reward Modeling in RL
Integration and Theoretical Foundations
Challenges and Strategies
Primary Functions
• Credit assignment problem
• Inefficiencies in preference learning
• Risk-based utility functions
• Assess value of actions
• Transform human feedback into learning signals
• Design effective reward functions
Future Directions in Reward Modeling
Expressive Feedback Mechanisms
Optimization and Efficiency
Enhancing AI Alignment and Adaptability
• Broader generalization techniques
• Comparison with other fine-tuning methods
• KL and uncertainty regularization
• Adaptive selection of calibration policies
Innovative Approaches to Reward Modeling
Regret-based Preference Models
Framework and Methodologies
Statistical and Theoretical Insights
• Quantify human preferences
• Ensure identifiability of reward functions
• Categorization of human feedback
• Stability and generalization challenges
• Minimax statistical bounds
• Impact on L alignment
• Sequence-to-sequence reward modeling
• Robust evaluation frameworks

Figure 2: This figure illustrates the hierarchical structure of key concepts in reward modeling within reinforcement learning, highlighting the roles, challenges, innovative approaches, and future directions in the field. It categorizes the main ideas into five primary sections, each with subcategories and detailed points, emphasizing the significance of reward function design and its impact on RL algorithms.

Designing effective reward functions is crucial as they directly influence agents' learning trajectories and decision-making capabilities.

A primary challenge in reward modeling is the credit assignment problem, which involves identifying actions leading to success or failure, especially in environments with sparse or delayed rewards. This is compounded by inefficiencies in existing preference learning methods, often hindered by complexities in reward modeling and KL-regularization [29]. Innovative strategies, such as risk-based utility functions, have been proposed to enhance decision-making by considering both expected returns and associated risks [2].

Integrating human feedback into reward modeling is crucial for aligning RL systems with human values, achieved by crafting reward functions that encapsulate human preferences across tasks [4]. Theoretical foundations of reward generalization in RLHF contribute to understanding reward dynamics from macro and micro perspectives, aiding in developing resilient reward models [30]. Game-theoretic approaches create a unified framework for generalization in RL, incorporating adversarial manipulations of task distributions to refine reward modeling [17]. Critiques of models assuming human preferences based solely on partial returns have led to new models grounded in regret, offering a nuanced perspective on reward dynamics [31].

The effectiveness of RLHF methods in aligning large language and vision-language models with human preferences underscores the critical role of reward modeling in RL [32]. By addressing challenges like credit assignment and employing innovative methodologies, researchers can enhance RL agents' capabilities to perform complex tasks while ensuring alignment with human values.

AI-generated, for reference only.

3.2 Challenges in Designing Reward Functions

Designing effective reward functions in Reinforcement Learning (RL) presents challenges that significantly impact learning processes and algorithm performance. Sparse reward signals lead to inefficient learning, exacerbated by instability in simultaneous training of actor and critic components in standard actor-critic methods [33]. This issue is particularly pronounced in environments with large action spaces, where existing benchmarks often fail to address RL instability and lack open-source resources for training language models with human feedback [4].

Figure 3 illustrates the primary challenges in designing reward functions for reinforcement learning, categorized into sparse rewards, human feedback, and offline reinforcement learning issues. Each category highlights specific problems such as inefficient learning, misrepresentation of preferences, and overestimation issues, as well as their associated references.

Deriving well-defined reward functions from human feedback is challenging, as current methods may misrepresent human preferences by favoring suboptimal actions due to higher partial returns [31]. Incorrect and ambiguous preference pairs in datasets exacerbate this issue, hindering accurate capture of human intent and resulting in poor generalization when trained on specific distributions [34].

In offline RL, classical off-policy algorithms struggle to estimate the value of out-of-distribution actions effectively, leading to inaccurate high value estimates [21]. This highlights the difficulty in designing reward functions that generalize well across diverse scenarios, particularly in dynamic real-world environments.

Applying primal-dual methods to Safe Reinforcement Learning (SRL) problems reveals interdependence of primal Learning Rate (LR) and Lagrangian Multiplier (LM) parameters, complicating the optimization process and design of reward functions ensuring safety and robustness [22].

Addressing challenges in robust and adaptable reward modeling techniques in RL is essential for advancing the field and enhancing performance of large language models (LLMs) across diverse applications. Complexity in developing effective reward models, especially in the context of RLHF and its susceptibility to noise, necessitates innovative approaches to improve robustness and adaptability. Recent research emphasizes designing reward functions that consider noise and employing methods like Direct Preference Optimization (DPO) to better align LLM outputs with human preferences. By tackling these challenges, researchers can facilitate systematic advancements in RL-enhanced LLMs, ultimately leading to improved alignment and performance in real-world scenarios [35, 10, 36].

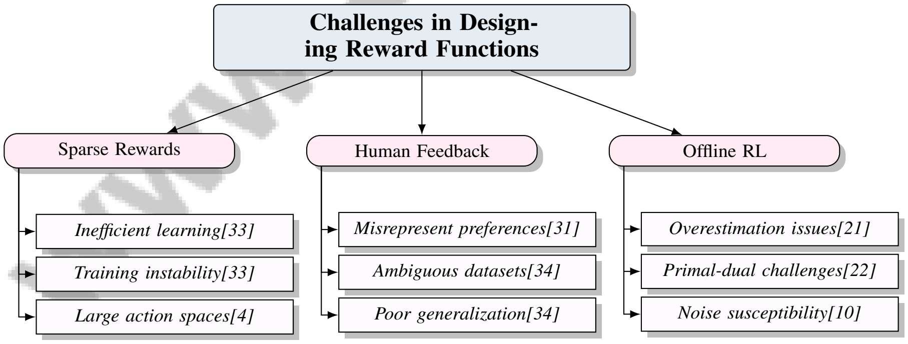


Figure 3: This figure illustrates the primary challenges in designing reward functions for reinforcement learning, categorized into sparse rewards, human feedback, and offline reinforcement learning issues. Each category highlights specific problems such as inefficient learning, misrepresentation of preferences, and overestimation issues, as well as their associated references.

3.3 Impact of Reward Function Design on RL Algorithms

The design of reward functions is crucial in shaping the performance and outcomes of Reinforcement Learning (RL) algorithms. A primary challenge is the reduction of policy entropy over time, leading to suboptimal exploration and learning. This necessitates adaptive reward function designs that balance exploration and exploitation, ensuring agents do not prematurely converge to suboptimal policies [37].

In Proximal Policy Optimization (PPO) algorithms, the static nature of the entropy coefficient limits the algorithm's ability to adaptively respond to agent performance during training, resulting in suboptimal exploration and learning outcomes [38]. Integrating adaptive mechanisms within reward functions is essential for enhancing the flexibility and responsiveness of RL algorithms to dynamic environments.

Balancing competing objectives of policy and value function training is another critical aspect of reward function design. When using a shared network, interference between these objectives can lead to suboptimal performance, underscoring the importance of carefully crafted reward functions that harmonize these competing goals [39]. Moreover, the delay between actions and their eventual effects complicates accurate assessment of each action's contribution to the final outcome, necessitating reward functions that effectively handle temporal credit assignment issues [11].

Innovative approaches like Discovered Policy Optimization (DPO) have shown significant advantages over traditional RL methods, requiring less hyperparameter tuning and achieving superior performance across diverse settings. These approaches highlight the potential for reward functions that are both theoretically sound and practically effective, facilitating more robust learning outcomes [40]. DPO agents have achieved significantly higher success rates compared to PPO agents, emphasizing the impact of well-designed reward functions on learning efficiency and success rates [41].

Theoretical advancements, such as those demonstrated by Optimistic Proximal Policy Optimization (OPPO+), offer improved regret guarantees for both stochastic and adversarial linear Markov Decision Processes (MDPs). These advancements illustrate the potential for reward functions to enhance the robustness and adaptability of RL algorithms in complex and uncertain environments [42].

The design of reward functions is crucial in shaping the performance and efficiency of RL algorithms, directly influencing the alignment of models with human preferences in RL from Human Feedback (RLHF) and impacting the robustness of agents in noisy environments, where perturbed rewards can lead to significant variations in learning outcomes [43, 36]. By addressing the challenges associated with reward function design, researchers can develop more robust and adaptable RL systems capable of navigating complex environments and achieving superior performance outcomes.

3.4 Innovative Approaches to Reward Modeling

Recent advancements in reward modeling within Reinforcement Learning (RL) have introduced innovative methodologies that significantly enhance the design and application of reward functions. A notable development is the integration of regret-based preference models, which quantify human preferences based on the regret of trajectory segments. This approach ensures the identifiability of reward functions, providing a more nuanced understanding of human preferences and their impact on RL systems [31].

The exploration of human feedback in RL has been enriched by a conceptual framework categorizing feedback into nine dimensions, encompassing human-centered, interface-centered, and model-centered aspects. This structured analysis offers a comprehensive way to evaluate existing research and identify areas for improvement in reward modeling [44].

Innovative methodologies have also been proposed to address stability and generalization challenges in reward models. Techniques enhancing these aspects have shown significant improvements in reward model performance, which are crucial for reliable and robust RL systems [34]. Additionally, the introduction of minimax statistical bounds for both Reinforcement Learning from Human Feedback (RLHF) and Discovered Policy Optimization (DPO) provides clearer insights into their performance across various settings, highlighting their effectiveness and limitations [45].

These advancements underscore the essential role of reward modeling in RL, particularly concerning reinforcement learning from human feedback (RLHF). They emphasize ongoing research efforts aimed at refining methodologies through which RL systems interpret and respond to environmental feedback, addressing challenges such as incorrect generalization, model misspecification, and feedback sparsity. This focus on reward models is crucial for improving the effectiveness of large language models (LLMs) and ensuring they align more closely with human preferences, thereby enhancing overall performance and usability [26, 10, 13, 12]. By addressing existing challenges and leveraging novel methodologies, these developments pave the way for more robust and effective learning outcomes in diverse applications.

3.5 Future Directions in Reward Modeling

The future of reward modeling in Reinforcement Learning (RL) is set for significant advancements, particularly in enhancing AI systems' alignment with human values and improving adaptability across diverse applications. One promising avenue is the application of sequence-to-sequence reward modeling (seq2seq RM) to larger models and a broader range of tasks, including those beyond preference datasets, to evaluate its effectiveness and scalability [46]. This approach could broaden reward models' applicability and enhance their generalization capabilities.

Future research should also focus on developing more robust evaluation frameworks for reward models, emphasizing synthetic data generation and enhancing user engagement during the training process [26]. This can lead to more reliable and user-aligned reward systems. Additionally, balancing KL and uncertainty regularization represents a critical area for optimizing performance, ensuring reward models maintain stability and robustness across various scenarios [47].

The adaptive selection of calibration policies and refinements in frameworks like Value-Incentivized Preference Optimization (VPO) are essential for enhancing applicability across different RL contexts [48]. Further refinements in algorithms for better performance in sparse feedback scenarios and exploring applications of Partially Observed Reinforcement Learning (PORRL) can contribute to more efficient and effective reward modeling [49].

Exploring further optimizations of methods like the Actor-Critic with Dynamic Baselines (ACDB), including adaptations for various task types and incorporating additional contextual information, could lead to improvements in sample efficiency and task-specific performance [50]. The application of ensemble methods in online RLHF setups and with larger-scale language models also holds promise for validating findings across different contexts and enhancing the robustness of reward models [51].

Future research should delve into additional feedback types, improve the calibration of human feedback models, and investigate the impact of explainability on feedback quality [52]. Developing more expressive feedback mechanisms and exploring interdisciplinary approaches will be crucial for refining the taxonomy of human feedback and accommodating diverse contexts [44]. Broader generalization techniques and comparisons of Parameter-Efficient Reinforcement Learning from Human Feedback (PE-RLHF) with other parameter-efficient fine-tuning methods are also vital areas for exploration [32].

The evolving landscape of reward modeling research in reinforcement learning (RL) emphasizes its capacity to catalyze significant advancements, particularly concerning large language models (LLMs). As demonstrated by the complexities of implementing RL-enhanced LLMs, researchers face challenges in developing effective algorithms and reward modeling strategies. Current studies highlight the importance of techniques like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) in aligning LLM outputs with human expectations. Moreover, innovative approaches, such as regularizing hidden states, have shown promise in improving reward models' generalization capabilities, addressing issues like reward over-optimization. Collectively, these future directions illustrate the dynamic nature of this field and point to vital areas for further exploration and improvement, essential for driving the next wave of breakthroughs in RL [10, 53]. By addressing current challenges and leveraging novel methodologies, researchers can develop more robust, adaptable, and human-aligned AI systems.

4 Policy Gradient Methods

To comprehend the advancements in policy optimization, it is crucial to examine the methodologies emerging within Policy Gradient Methods. These methods have been instrumental in refining training processes and improving the performance of Reinforcement Learning (RL) systems. Table 1 presents a detailed categorization of advanced policy optimization methods, illustrating the diverse strategies and innovations enhancing policy gradient methods in Reinforcement Learning. Additionally, Table 2 presents a comprehensive comparison of advanced policy optimization methods, elucidating their unique strategies and impact on reinforcement learning advancements. The subsequent subsection explores advanced policy optimization methods, highlighting key innovations and their contributions to the field.

Category	Feature	Method
Iterative Policy Update Mechanisms	Clipping and Optimization Techniques	Pb-PPO[54], PPO[55], TRGPPO[56], PPO[6]
	Exploration and Sampling Strategies	SAPG[57], PPO-UE[58], RISE[59]
	Feedback and Trajectory Methods	RPO[60], MDPPO[61], ZPG[62], DLADD[18]
	Gradient and Stability Enhancements	ROBUST-PPO-NOCLIP[63], PPO-B[64], GI-PPO[20]
Comparison with Other Policy Gradient Methods	Adaptive Techniques	APD[22]

Table 1: This table provides a comprehensive overview of advanced policy optimization methods categorized under Iterative Policy Update Mechanisms and their comparison with other Policy Gradient Methods. It highlights various features and methods, including Clipping and Optimization Techniques, Exploration and Sampling Strategies, and Adaptive Techniques, with references to key studies that contribute to the field of Reinforcement Learning.

4.1 Advanced Methods for Policy Optimization

Advanced policy optimization methods in RL have significantly enhanced the robustness and efficiency of training algorithms. Proximal Policy Optimization (PPO) is a cornerstone method, using a surrogate objective function to optimize policies while constraining updates for stable improvements. This method is further refined in Proximal Policy Optimization with Uncertainty Exploration (PPO-UE), which dynamically adjusts the exploration-exploitation balance based on uncertainty, improving convergence speed and performance, especially in Roboschool continuous control tasks [65, 58, 38, 66].

The hierarchical structure of these advanced policy optimization methods is illustrated in Figure 4, categorizing them into three main groups: Proximal Policy Optimization, Innovative Gradient Methods, and Meta-Learning and Safety. Each category encompasses specific techniques that highlight their contributions to enhancing policy optimization performance.

The Split and Aggregate Policy Gradients (SAPG) method innovatively divides complex environments into manageable blocks, training diverse policies tailored to each block. This technique uses off-policy updates to aggregate data, enhancing performance in challenging environments compared to traditional methods like PPO [57, 39]. This highlights the importance of leveraging diverse data sources to boost policy performance.

Gradient Informed Proximal Policy Optimization (GI-PPO) integrates analytical gradients into the PPO framework, introducing an -policy for adaptive gradient influence, improving performance across applications like function optimization and traffic control [67, 68, 20]. Similarly, Behavior Proximal Policy Optimization (BPPO) improves behavior policy while ensuring monotonic policy improvement, particularly in offline RL settings.

The decision-aware actor-critic framework enhances sample efficiency by jointly optimizing the actor and critic, aligning their objectives to maximize cumulative rewards [69, 70, 71, 39, 72]. This synchronized training approach significantly boosts policy optimization.

Innovative methodologies like Pairwise Proximal Policy Optimization (P3O) enhance optimization by adopting trajectory-wise approaches, while Predictive Processing Proximal Policy Optimization (P4O) uses predictive processing to minimize surprise and improve cumulative rewards, outperforming traditional methods in complex tasks [15, 11, 9].

The Learnt Policy Optimisation (LPO) and Discovered Policy Optimisation (DPO) frameworks exemplify meta-learning's potential in policy optimization, achieving state-of-the-art performance in Brax environments [37, 60, 40, 73, 74]. These methods underscore the role of adaptive learning strategies in refining policy optimization processes.

Moreover, the Adaptive Primal-Dual (APD) method enhances Safe Reinforcement Learning (SRL) performance by incorporating adaptive learning rates based on Lagrangian multipliers [22].

Advanced RL methods, particularly those enhancing large language models (LLMs) like DeepSeek-R1, illustrate the field's dynamic evolution by integrating strategies such as RLHF and DPO. These approaches leverage human feedback to align LLM outputs with user expectations, addressing challenges like reward modeling and optimization intricacies [10, 12]. By leveraging these novel methodologies, researchers continue to push RL systems' capabilities, paving the way for more sophisticated learning outcomes.

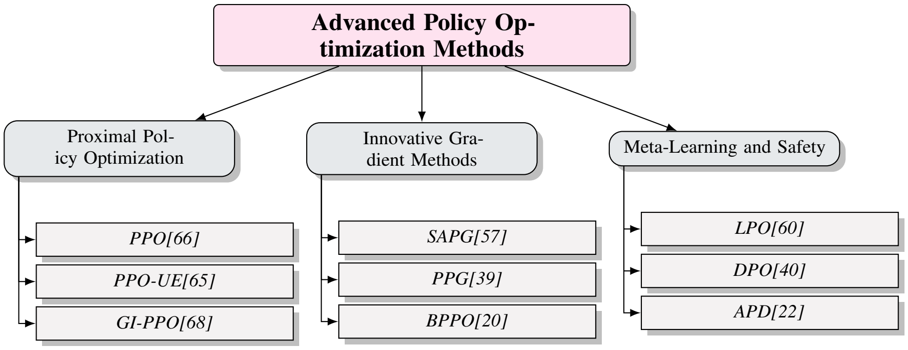


Figure 4: This figure illustrates the hierarchical structure of advanced policy optimization methods in reinforcement learning, categorizing them into Proximal Policy Optimization, Innovative Gradient Methods, and Meta-Learning and Safety. Each category includes specific techniques, highlighting their contributions to enhancing policy optimization performance.

4.2 Iterative Policy Update Mechanisms

Iterative policy update mechanisms are vital in policy gradient methods within RL, enabling continuous policy improvement to optimize expected rewards. Natural Policy Gradient (NPG) uses the Fisher information matrix to ensure stable and efficient updates, addressing issues like heavy-tailed gradients in algorithms like PPO [39, 63]. Enhancements like NPG-A, which incorporates adaptive step sizes, further boost convergence rates and robustness.

PPO and its variants are prominent for their stability and efficiency, maximizing a clipped surrogate objective function to balance exploration and exploitation [6]. Proximal Policy Gradient (PPG) refines policies by adjusting action probabilities using a modified advantage function, while Trust Region Guided Proximal Policy Optimization (TRGPPO) enhances exploration by adapting the clipping range within a trust region [56]. Pb-PPO employs a bi-level optimization framework to dynamically adjust the clipping bound during training [54].

Innovative methodologies like ROBUST-PPO-NOCLIP optimize policy using a robust estimator, mitigating the impact of heavy-tailed gradients [63]. PPO-B uses a logarithmic barrier function to ensure updates remain within the feasible region, preventing divergence and improving gradient estimation [64].

SAPG addresses data duplication by ensuring diverse action sampling across multiple policies, enhancing learning in high-dimensional state spaces [57]. MDPPO mixes complete and auxiliary trajectories from multiple policies, providing a robust training dataset [61]. Intrinsically motivated reinforcement learning, such as RISE, quantifies state novelty and provides intrinsic rewards to encourage exploration [59].

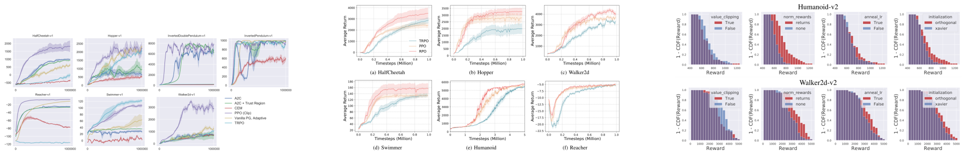
Zeroth-Order Policy Gradient (ZPG) estimates policy gradients directly from human feedback, utilizing a zeroth-order optimization approach [62]. This method exemplifies leveraging human feedback in iterative policy updates to enhance adaptability and effectiveness.

GI-PPO introduces an adaptive -policy to adjust analytical gradients' influence based on variance and bias, improving upon traditional PPO methods [20]. The method dynamically adjusts to new data inputs, enhancing predictive capabilities [18]. PPO-UE employs self-adaptive exploration strategies to improve training stability [58].

Overall, these iterative mechanisms highlight policy gradient methods' dynamic and adaptive nature, underscoring their capacity to navigate diverse RL environments. By incorporating cutting-edge strategies, current RL research—especially regarding LLMs—significantly enhances capabilities. Techniques like RLHF and DPO improve LLM output alignment with human preferences, addressing challenges like feedback sparsity and model misspecification, paving the way for sophisticated learning outcomes [10, 13, 12].

As shown in Figure 5, policy gradient methods and iterative policy update mechanisms are crucial for optimizing decision-making in RL across various environments. These examples illustrate their efficacy in different contexts. The first image shows learning curves of various RL algorithms,

AI-generated, for reference only.



(a) Comparison of Learning Curves for Different RL Algorithms on Various Gym Environments[55]

(b) Comparison of Average Returns Across Different RL Algorithms for Various Tasks[60]

(c) Humanoid-v2 and Walker2d-v2: Comparison of Reward Distribution and CDF[75]

Figure 5: Examples of Iterative Policy Update Mechanisms

such as A2C and PPO, interacting with Gym environments, highlighting cumulative rewards as performance indicators. The second image examines average returns achieved by TRPO, PPO, and RPO across tasks like HalfCheetah and Humanoid, emphasizing efficiency over millions of timesteps. Lastly, the third image provides a detailed look at reward distribution and CDF for Humanoid-v2 and Walker2d-v2 environments, offering insights into reward variability and distribution. Together, these figures encapsulate iterative policy updates' diverse approaches and outcomes in RL, demonstrating their impact on agent performance across tasks and environments [55, 60, 75].

4.3 Comparison with Other Policy Gradient Methods

Policy gradient methods are integral to RL, providing a framework for directly optimizing policies by estimating expected reward gradients. Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO) are popular for balancing stability and efficiency. Recent advancements in language model training, like Fine-Grained RLHF and SuperHF, enhance traditional approaches by improving performance and stability, allowing nuanced feedback on errors like factual inaccuracies, addressing safety, and alignment challenges in language models [25, 5].

Enhanced Actor-Critic with Maximum Entropy (EAPO) improves over baseline algorithms like PPO and TRPO by maintaining high entropy in policy distributions, promoting robust exploration and learning outcomes across tasks, as shown by superior performance in MuJoCo and Procgen environments [37, 76, 70, 77, 72]. This entropy maintenance prevents premature convergence and ensures comprehensive policy space exploration. Similarly, Natural Policy Gradient with Adaptive step sizes (NPG-A) surpasses traditional methods by achieving faster convergence and superior policy optimization.

Reinforcement Learning with Optimistic Optimization (RLOO) employs multiple samples to mitigate variance, enhancing performance and efficiency compared to traditional methods like PPO, which suffer from high variance in surrogate objectives due to importance sampling [77, 73, 42, 65]. This underscores sample efficiency's importance in achieving optimal policy performance, particularly in high-dimensional environments. Dataset Reset Policy Optimization (DR-PO) outperforms PPO and DPO by generating high-quality outputs while maintaining computational efficiency, highlighting strategic dataset management's potential in enhancing policy optimization.

Policy Optimization with Model-based Explorations (POME) enhances PPO by integrating model-based explorations, improving sample efficiency and learning effectiveness. This approach combines model-free Monte Carlo sampling with a learned transition model, optimizing exploration strategies and enhancing decision-making in complex environments, as shown by superior performance in 33 out of 49 Atari games [71, 78]. However, challenges persist with softmax policy gradient methods, which can exhibit exponential iteration complexity, requiring careful application in large state spaces and long horizons.

Despite advancements, fundamental challenges remain. PPO and TRPO differences are often overshadowed by implementation optimizations' impact on performance; research indicates code-level enhancements significantly affect agent behavior, with PPO's performance gains over TRPO attributed to these optimizations, complicating performance evaluation in deep RL [79, 75, 80]. Existing PPO clipping mechanisms do not adequately restrict policy updates, allowing large detrimental changes, highlighting the need for refined mechanisms for stable updates.

Vanilla Policy Gradient (VPG) has demonstrated superior learning in certain contexts, often achieving higher average scores than PPO, indicating simpler algorithms can outperform complex methods in specific scenarios [57, 68, 20]. Moreover, Behavior Proximal Policy Optimization (BPPO) demonstrates superior performance in offline RL tasks by achieving monotonic policy improvement without additional constraints.

The Adaptive Primal-Dual (APD) method exemplifies policy optimization advancements, significantly outperforming traditional primal-dual methods with constant learning rates, demonstrating improved stability and performance [22]. This method's dynamic learning rate adjustment based on environmental feedback underscores adaptability's importance in RL algorithms.

Overall, while traditional methods like PPO and TRPO have laid policy optimization groundwork in RL, emerging methods like EAPO, NPG-A, and DR-PO offer promising alternatives addressing limitations and enhancing learning performance. By implementing advanced techniques like Successive Policy Re-weighting (SPR) and enhancing exploration through intrinsic exploration modules (IEM-PPO), researchers address key RL challenges. These strategies improve sample efficiency and computational feasibility for LLMs, facilitating better alignment with human values, driving significant advancements in RL-based policy optimization [79, 81, 71, 66, 82].

Feature	Proximal Policy Optimization (PPO)	Proximal Policy Optimization with Uncertainty Exploration (PPO-UE)	Split and Aggregate Policy Gradients (SAPG)
Optimization Approach	Surrogate Objective Function	Dynamic Exploration Balance	Environment Decomposition
Stability Mechanism	Constrained Updates	Uncertainty-based Adaptation	Off-policy Updates
Application Domain	General RL Tasks	Roboschool Tasks	Complex Environments

Table 2: This table provides a comparative analysis of three advanced policy optimization methods: Proximal Policy Optimization (PPO), Proximal Policy Optimization with Uncertainty Exploration (PPO-UE), and Split and Aggregate Policy Gradients (SAPG). It highlights the distinctive optimization approaches, stability mechanisms, and application domains of each method, illustrating their respective contributions to enhancing reinforcement learning performance.

5 Proximal Policy Optimization (PPO) and Generalized Proximal Policy Optimization (GRPO)

Proximal Policy Optimization (PPO) and Generalized Proximal Policy Optimization (GRPO) are pivotal in optimizing complex decision-making processes across various domains due to their versatility and efficacy. The following subsection elucidates specific applications of these methodologies, highlighting their robustness and adaptability in real-world scenarios.

5.1 Applications and Use Cases

PPO and GRPO have been successfully applied in diverse fields owing to their robustness in high-dimensional policy spaces. The Proximal Policy Gradient Arborescence (PPGA) enhances exploration and policy development in complex environments through vectorized implementations [57]. In continuous action spaces, PPO's adaptability is evident through its superior performance in safety-critical tasks like multiclass queueing and ride-hailing optimization, surpassing Clipped Objective Policy Gradient (COPG) and Trust Region Policy Optimization (TRPO) [6, 83].

PPO also plays a significant role in refining large-scale language models, working with Direct Preference Optimization (DPO) to align learning processes with human preferences through token-wise reward functions [73]. Optimistic Proximal Policy Optimization (OPPO+) excels in adversarial settings, particularly in episodic adversarial linear Markov Decision Processes (MDPs), demonstrating its potential in challenging environments [42].

In autonomous driving, PPO has been integrated into HCPI-RL planners, outperforming rule-based and traditional PPO planners in emergent and daily cruising scenarios [1]. It has also demonstrated effectiveness in high-fidelity simulations, such as with the Iris quadcopter, where it outperforms baseline PID controllers [3].

PPO's versatility is further illustrated in multi-agent contexts within the Unity platform, where experiments using the Unity ML-Agents Toolkit show its efficacy [61]. Variants like Gradient-Informed Proximal Policy Optimization (GI-PPO) and PPO-UE have proven effective in diverse environments, enhancing robustness in handling training uncertainties [20, 58].

In conversational AI, PPO has been applied to chit-chat chatbots, leveraging datasets like OpenSubtitles to outperform baseline methods such as REINFORCE and SeqGAN [84]. The diverse applications of PPO and its variants, including Appraisal-Guided Proximal Policy Optimization (AG-PPO) and Reinforced Token Optimization (RTO), underscore their potential to enhance decision-making processes, achieving state-of-the-art results in complex tasks such as code generation and dialogue systems [85, 86, 73, 83, 80].

As shown in Figure 6, this figure illustrates the diverse applications and use cases of PPO and GRPO, highlighting their adaptability in high-dimensional policy spaces, language models, AI, and autonomous systems. Each category demonstrates the algorithms' effectiveness in enhancing exploration, refining large-scale models, and improving autonomous system performance. The figures compare these algorithms' performances across diverse tasks, emphasizing their utility in refining model performance and enhancing computational efficiency.

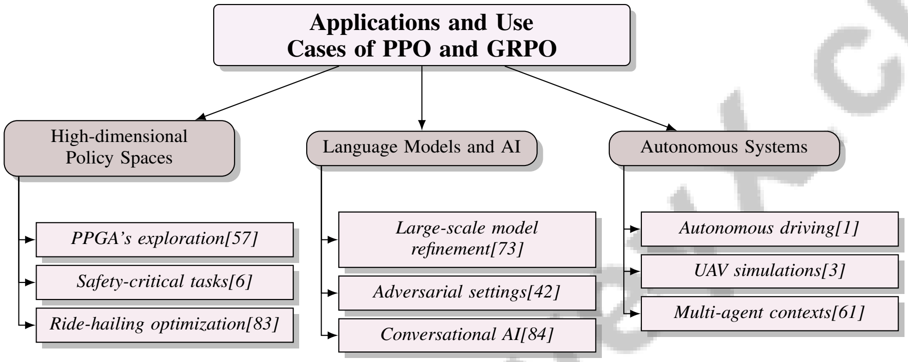


Figure 6: This figure illustrates the diverse applications and use cases of PPO and GRPO, highlighting their adaptability in high-dimensional policy spaces, language models, AI, and autonomous systems. Each category demonstrates the algorithms' effectiveness in enhancing exploration, refining large-scale models, and improving autonomous system performance.

5.2 Challenges and Future Directions

Despite their strengths, PPO and GRPO face challenges that constrain their broader applicability. A key issue is their sensitivity to hyperparameter tuning, which can be resource-intensive, as seen in PPO-dynamic's performance enhancements [84]. The reliance on deterministic environments, as evidenced by methods like Predictive Processing Proximal Policy Optimization (P4O), limits their robustness in stochastic real-world scenarios [9]. Similarly, Behavior Proximal Policy Optimization (BPPO) depends heavily on behavior policy quality, which can undermine performance if suboptimal [21].

The computational complexity of enhanced exploration mechanisms in PPO presents scalability challenges, particularly in multi-agent and high-dimensional environments [58]. Estimating hyperparameters in methods like Adaptive Primal-Dual (APD) remains problematic, necessitating further research for optimization in complex RL settings [22].

Future research should focus on developing adaptive hyperparameter tuning methods to enhance PPO and GRPO's robustness across diverse applications. Extending frameworks like PPO-UE to multi-agent scenarios and more complex environments offers a promising direction, potentially increasing these methods' robustness in real-world contexts [58]. Additionally, integrating sophisticated exploration strategies and refining policy aggregation methods could further enhance these algorithms' adaptability and effectiveness [57].

By addressing these challenges and exploring innovative future directions, PPO and GRPO can reinforce their roles as leading methodologies in reinforcement learning, offering tailored solutions for complex decision-making tasks across various domains, including large language model alignment and diverse demographic preferences [73, 86, 87, 80].

6 Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) is crucial for aligning AI systems with human values and preferences, yet it faces significant challenges in reward modeling and integrating human feedback. These challenges, including incorrect generalization, model misspecification, and feedback scarcity, impact large language models' (LLMs) ability to embody human values like helpfulness, harmlessness, and honesty [88, 23, 12].

6.1 Challenges in Reward Modeling and Feedback Integration

The integration of reward modeling with human feedback in RLHF frameworks is impeded by the high cost and time required for acquiring quality human preference labels, limiting scalability. Fragility in optimization processes, due to noise in reward models, can mislead policy learning, resulting in suboptimal outcomes. Ambiguity in preference data, often stemming from low annotator agreement, complicates generalization across distributions [34].

Bias in AI outputs arises from the lack of diversity in human feedback, undermining RLHF systems' reliability. The absence of standardized feedback classification and challenges in expressing feedback effectively hinder communication between humans and AI agents [44]. The Bradley-Terry model, among others, fails to capture the full range of human preferences, necessitating more sophisticated models [89].

The resource-intensive nature of current RL methods, which require loading multiple models simultaneously, limits practical applicability [90]. The lack of comprehensive statistical analyses comparing RLHF with Direct Preference Optimization (DPO) further complicates understanding the conditions under which one method may outperform the other [45]. Innovative approaches, such as regret-based preference models, offer a more accurate representation of human preferences [31]. Benchmarks like Parameter-Efficient Reinforcement Learning from Human Feedback (PE-RLHF) focus on aligning large models with human preferences while minimizing computational costs [32]. Addressing these challenges can enhance reward modeling integration with human feedback, leading to more robust RL systems.

6.2 Scalability and Efficiency in RLHF

Scalability and efficiency are vital for RLHF's implementation in large-scale AI systems, crucial for aligning AI models with human preferences. Ensemble-based conservative optimization objectives, such as worst-case optimization (WCO) and uncertainty-weighted optimization (UWO), provide robust frameworks accommodating variability in human feedback [51].

Developing personalized reward models through representation learning and clustering enhances scalability by aggregating preferences from multiple human labelers via probabilistic opinion feedback, improving efficiency [91]. The Reinforced Token Optimization (RTO) approach, employing token-wise rewards, improves sample efficiency and aligns AI models with human preferences, underscoring fine-grained reward structures' importance [73].

Active-query-based methods reduce the number of queries needed for aligning LLMs with human preferences, streamlining the RLHF process [92]. RRLHF, recognizing imperfections in reward models and utilizing contrastive rewards, enhances robustness and scalability by addressing noise and ambiguity in human feedback [35].

The integration of R3M, an 1-regularized maximum likelihood estimation approach, tackles issues of corrupted preference labels, improving reward models' robustness and scalability in RLHF [93]. The MinorDPO approach facilitates higher learning rates without destabilizing the model, suggesting scalability and efficiency improvements through aggressive learning strategies [94]. Future research should prioritize scaling RLHF to larger datasets and enhancing computational efficiency, as indicated by ongoing DPO framework improvements [33]. RLAIF presents a scalable solution for aligning language models, reducing costs associated with gathering human preference labels [95]. Addressing these scalability and efficiency considerations can evolve RLHF into a robust framework for aligning AI systems with human values.

6.3 Integration of Human Feedback in Complex Environments

Integrating human feedback in complex RL environments is essential for aligning AI systems with human values in dynamic settings. The complexity and variability of tasks necessitate advanced methods for effective human feedback incorporation. The Nash Learning from Human Feedback (NLHF) method introduces a pairwise preference model establishing a Nash equilibrium for preferred responses, facilitating better LLM alignment with human preferences [89]. This approach accounts for strategic interactions between agents and human evaluators, enhancing learning robustness.

As illustrated in Figure 7, the integration of human feedback in complex RL environments highlights key methodologies such as Nash Learning for strategic interactions, Parameter-Efficient RLHF for computational efficiency, and strategies to enhance feedback diversity and quality. The Parameter-Efficient Reinforcement Learning from Human Feedback (PE-RLHF) setup significantly reduces the computational burden of traditional RLHF frameworks, making it more accessible for aligning large models with human feedback [32]. This efficiency is particularly advantageous in complex environments with limited computational resources, allowing broader application and experimentation with human feedback integration techniques.

To mitigate challenges related to feedback diversity and quality, employing pluralistic panels for evaluation is recommended. This strategy fosters inclusivity and representation in feedback, reducing biases and enhancing RLHF processes' robustness. An ethical framework emphasizes inclusivity in developing adaptive feedback mechanisms, addressing model drift and human preferences' dynamic nature. By advocating a pluralistic approach, the framework ensures AI systems are responsive to diverse human needs and values, improving efficacy in various applications [26, 91, 27, 96].

Voting mechanisms across multiple reward models can improve human feedback integration by allowing nuanced assessments of preference strength, capturing diverse aspects of user evaluations like factual accuracy and relevance, as demonstrated by the Fine-Grained RLHF framework. This enhances feedback granularity and enables language model behavior customization based on specific error types [25, 26]. Combined with contrastive learning and meta-learning techniques, these methodologies improve reward models' generalization and robustness, facilitating reliable training across various distributions. Innovative approaches like the SPR method, enabling policy optimization with minimal memory usage, significantly enhance scalability and efficiency, addressing traditional RLHF limitations.

Integrating human feedback in complex RL environments requires a comprehensive strategy incorporating innovative methodologies and practical approaches, as evidenced by critical RLHF analyses for LLMs. This involves understanding reward model intricacies, addressing limitations like incorrect generalization and feedback sparsity, and employing a taxonomy of feedback types bridging human-centered, interface-centered, and model-centered perspectives. Identifying quality metrics and design requirements for effective human feedback systems enhances agent learning capabilities and user interaction, paving the way for interdisciplinary collaboration to optimize RL applications [12, 44]. By tackling existing challenges and leveraging novel insights, researchers can develop robust and adaptable RL systems capable of navigating complex decision-making tasks.

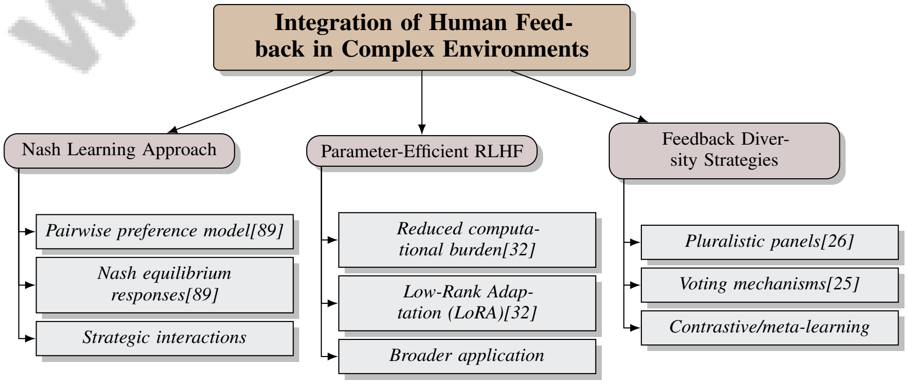

Integration of Human Feedback in Complex Environments
Nash Learning Approach
Parameter-Efficient RLHF
Feedback Diversity Strategies
Pairwise preference model[89]
Nash equilibrium responses[89]
Strategic interactions
Reduced computational burden[32]
Low-Rank Adaptation (LoRA)[32]
Broader application
Pluralistic panels[26]
Voting mechanisms[25]
Contrastive/meta-learning

Figure 7: This figure illustrates the integration of human feedback in complex RL environments, highlighting key methodologies such as Nash Learning for strategic interactions, Parameter-Efficient RLHF for computational efficiency, and strategies to enhance feedback diversity and quality.

7 Large-Scale Language Models in Reinforcement Learning

To appreciate the transformative impact of large-scale language models (LLMs) on reinforcement learning (RL), it is crucial to examine their contributions within these systems. The following subsections explore how LLMs enhance RL frameworks by processing complex linguistic inputs and facilitating effective decision-making, highlighting the intersection of LLMs and RL.

7.1 Role of Large-Scale Language Models in Reinforcement Learning

LLMs have become central to advancing RL frameworks by processing and generating human-like language, thus improving interactions between RL systems and users. Their integration, through methods like reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), enables nuanced, context-aware decision-making, aligning AI outputs with human preferences and values. This is crucial for refining model behavior and addressing issues such as toxicity and hallucinations [10, 12].

Incorporating LLMs refines policy optimization via methodologies like Multi-Preference Optimization (MPO), which uses a natural actor-critic framework to optimize model responses based on user feedback, enhancing adaptability in dynamic environments [97]. Approaches like Contextualized Critiques with Constrained Preference Optimization (C3PO) leverage synthetic preference data from high-level verbal feedback to fine-tune language models, showcasing LLMs' capacity to incorporate diverse human feedback and improve interaction quality [98].

The integration of LLMs into RL frameworks signifies a major advancement, enhancing the processing of complex linguistic inputs and generating contextually appropriate responses. By utilizing techniques such as RLHF, researchers can develop RL systems that align more closely with human values, effectively addressing challenges and improving AI solutions. This approach emphasizes explicit reward signals and advanced modeling strategies for understanding human expectations, enhancing AI behavior alignment with societal norms [10, 13, 27, 99, 12].

7.2 Integration Techniques for LLMs in RL

Integrating LLMs into RL systems is pivotal for advancing AI capabilities, enabling sophisticated interactions and decision-making processes. Preference-based optimization frameworks leverage LLMs' expressive power to interpret nuanced human feedback, aligning AI systems with human values by incorporating detailed linguistic inputs into the RL training process [97].

Techniques like C3PO exemplify LLM potential in RL by using synthetic preference data for fine-tuning language models, allowing RL systems to incorporate high-level verbal feedback for more contextually relevant policy adjustments [98]. Additionally, Multi-Preference Optimization (MPO) within a natural actor-critic framework iteratively optimizes model responses based on preference feedback, ensuring AI systems align with user expectations and effectively navigate complex conversational contexts [97].

Advancements in token-wise reward structures further facilitate LLM integration into RL frameworks by enhancing sample efficiency and aligning AI models with human preferences. Token-wise rewards allow RL systems to achieve a granular understanding of human feedback, leading to improved decision-making processes and accurate policy adjustments [73].

Integrating LLMs into RL systems represents a significant advancement, enhancing capabilities for processing complex linguistic inputs and generating contextually appropriate responses. Employing advanced integration techniques like RLHF enables the development of nuanced RL systems that better align with diverse human values and preferences, addressing complexities in human-AI interaction and enhancing AI solutions' effectiveness while mitigating issues like toxicity and hallucinations [26, 100, 27, 12].

7.3 Prompt Optimization and Evaluation

Effective prompt optimization and evaluation are critical for integrating LLMs with RL systems, significantly enhancing AI output alignment with human preferences and improving interaction quality. This complex integration requires sophisticated algorithms and reward modeling strategies, particularly in approaches like RLHF and DPO. Recent research emphasizes the need for systematic reviews and innovative techniques, such as Proximal Policy Optimization (PPO), to align LLM outputs with human expectations and enhance performance in formal language tasks [10, 99].

A promising approach to prompt optimization leverages reinforcement learning techniques, particularly RLHF, to iteratively refine prompts based on user interactions and feedback. This method employs advanced algorithms, like PPO, although simpler REINFORCE-style optimization variants may offer comparable performance with reduced computational costs. Focusing on aligning LLMs with human preferences enhances prompt effectiveness and addresses challenges related to reward model sensitivity and feedback sparsity, leading to robust learning outcomes [10, 101, 13, 28]. This feedback-driven approach allows RL systems to dynamically adjust prompts in response to evolving user inputs, enhancing AI flexibility and responsiveness.

Evaluating prompt effectiveness is essential to ensure that prompts used in RL with LLMs achieve desired outcomes. The evaluation process in RLHF encompasses both quantitative and qualitative assessments, measuring prompt performance against criteria such as coherence, relevance, factual accuracy, and user satisfaction. This comprehensive evaluation identifies aspects influencing user preferences, enhancing the model's ability to generate contextually appropriate and valuable responses [25, 102, 27, 12]. Systematic prompt evaluations enable researchers to identify areas for improvement and develop more effective strategies.

Integrating token-wise reward structures in RL frameworks facilitates a nuanced approach to prompt optimization by enabling fine-grained reward signal learning at the token level, as demonstrated in the Reinforced Token Optimization (RTO) algorithm. This method captures detailed preference data, improving LLM alignment and performance in interactive decision-making tasks, ultimately leading to more effective policy optimization compared to traditional sentence-level reward systems [10, 103, 104, 73]. Assigning rewards at the token level allows for finer control over the generation process, resulting in contextually relevant outputs and enhancing the sample efficiency of RL systems.

Prompt optimization and evaluation are vital for successfully integrating LLMs with RL systems. This process aligns LLMs with human behavior and expectations, involving complex algorithms and reward modeling strategies that enhance model performance. Effective prompt optimization aids in aligning LLM outputs with desired outcomes while addressing computational efficiency and resource management challenges in RL methods like RLHF. By refining prompts, researchers can navigate RL intricacies, leading to more robust and effective LLM applications across various tasks [10, 90, 13, 14, 99]. Leveraging advanced techniques and systematic evaluation processes enables the development of sophisticated AI models aligned with human values, capable of delivering high-quality interactions across diverse applications.

7.4 Benchmarking and Evaluation of LLMs in RL

Benchmark	Size	Domain	Task Format	Metric
RLHF-Bench[105]	200,000	Mathematics	Reasoning	Best-of-N, Reward
RLHF-Bench[106]	200,000	Mathematics	Reasoning	Best-of-N, PPO
Hydra-RLHF[107]	150,000	Language Modeling	Reinforcement Learning	PPL, RM accuracy
RLHF-LLM[108]	1,000	Mathematics	Question Answering	maj@1, pass@96
TL;DR[109]	18,065	Text Summarization	Summarization	ROUGE
BRIDGE[110]	155	Reinforcement Learning	Policy Optimization	Effective Horizon
RLHF-Bench[80]	200,000	Code Generation	Dialogue Generation	pass@k
PPO-D3[111]	57	Reinforcement Learning	Policy Optimization	Median Human Normalized Scores, Optimality Gap

Table 3: This table presents a diverse set of benchmarks employed for evaluating Large Language Models (LLMs) within Reinforcement Learning (RL) contexts. It details various benchmarks, their sizes, domains, task formats, and the metrics used to assess performance. These benchmarks are crucial for understanding the effectiveness of LLMs in complex decision-making tasks.

Benchmarking and evaluating LLMs within RL applications are crucial for understanding their performance in complex decision-making environments. A comprehensive benchmarking framework for LLMs should encompass diverse metrics that evaluate the alignment of generated outputs with human preferences, measure the efficiency of reinforcement learning processes (such as RLHF or DPO), and assess adaptability across tasks and contexts, ensuring effective responses to human instructions in real-world applications [90, 86, 10, 80]. Table 3 provides a comprehensive overview of the benchmarks utilized for assessing the performance of Large Language Models in Reinforcement Learning applications, highlighting the diversity in task domains and evaluation metrics.

One benchmarking approach involves utilizing synthetic preference data to evaluate models' effectiveness in incorporating human feedback. This method enables systematic comparisons of different models, identifying strengths and limitations in processing nuanced human inputs [98]. By leveraging synthetic data, researchers can simulate various scenarios and assess LLM robustness in handling diverse feedback types.

The evaluation process should include both quantitative and qualitative assessments for a holistic understanding of model performance. Quantitative metrics, such as reward alignment and task completion rates, provide objective measures of goal achievement, while qualitative assessments, including user satisfaction and interaction quality, offer insights into the subjective user experience with the models [97].

The integration of token-wise reward structures in RL frameworks provides a granular approach to evaluating LLMs. By assigning rewards at the token level, researchers gain insights into individual tokens' contributions to overall model performance, allowing for precise benchmarking and evaluation [73]. This approach enhances understanding of model behavior and informs the development of more effective training strategies.

Benchmarking and evaluating LLMs within RL applications require a multifaceted approach combining synthetic data, quantitative metrics, and qualitative assessments. Implementing comprehensive evaluation frameworks enables researchers to systematically analyze the capabilities and limitations of reinforcement learning-enhanced LLMs, such as those utilizing RLHF and other reward model techniques. This deeper understanding addresses the complexities inherent in optimizing these models and facilitates the development of robust and effective AI systems by identifying critical areas for improvement and guiding future research directions [10, 12].

8 Algorithmic Efficiency in Reinforcement Learning

In Reinforcement Learning (RL), achieving algorithmic efficiency is crucial as task and environment complexities increase. This efficiency involves optimizing computational speed, resource utilization, sample efficiency, and balancing exploration and exploitation strategies. Recent advancements, such as integrating diffusion models with Proximal Policy Optimization (PPO) and developing the Intrinsic Exploration Module (IEM-PPO), enhance exploration and sample efficiency in complex tasks. Algorithms like Successive Policy Re-weighting (SPR) optimize policy learning using both offline and online datasets, balancing resource constraints with high-quality data needs. These innovations collectively improve cumulative rewards, convergence speed, and strategy stability across various RL applications [82, 71, 66]. Each aspect contributes to a more efficient learning process, facilitating the development of robust RL systems capable of functioning in diverse and challenging contexts.

8.1 Optimization of Computational Resources

Optimizing computational resources in RL is essential for enhancing training efficiency and scalability, particularly in complex environments. Strategies like simplifying policy optimization by focusing on significant reward differences, as in REBEL, enable effective training with fewer resources [112]. Enhancements to PPO, including uncertainty estimation, guide exploration and reduce the risk of local optima convergence, stabilizing learning and optimizing resource use [66]. Masksembles improve out-of-distribution detection while maintaining competitive performance, contributing to efficient resource utilization [113].

Direct Preference Optimization (DPO) simplifies and stabilizes training for language models, effectively lowering computational barriers associated with traditional RL methods [114]. In environments with complex partial differential equations (PDEs), a multilevel RL framework reduces computational costs while enhancing optimal control policy learning [115]. Advanced RL algorithms integrated with specialized model architectures, as demonstrated by RLOR, enhance training speeds and performance [116]. PPO-B improves sampling efficiency and performance over original PPO in Atari and Mujoco environments, highlighting the potential for optimizing computational resources [64].

AI-generated, for reference only.

Predictive Processing Proximal Policy Optimization (P4O) offers improved learning efficiency and reduced computational requirements, outperforming state-of-the-art agents in challenging environments [9]. By leveraging innovative methodologies and refining existing techniques, researchers can enhance RL systems' effectiveness while minimizing resource consumption, emphasizing the potential for optimizing computational resources across diverse applications.

8.2 Enhancing Sample Efficiency

Enhancing sample efficiency in RL is crucial for accelerating learning processes and improving algorithm performance, particularly in environments with sparse rewards or complex dynamics. The Phasic Policy Gradient (PPG) method improves sample efficiency by decoupling policy and value function training, allowing tailored optimization strategies [39]. Fine-grained human feedback aligns RL systems with human preferences but introduces increased computational costs due to multiple evaluations [25]. Despite this challenge, detailed feedback enhances learning by providing precise guidance.

The IMPEC framework demonstrates improved sample efficiency and reduced reward confusion, achieving higher mean returns by leveraging intrinsic motivation to guide exploration [117]. HCPI-RL enhances sample efficiency by ensuring newly learned policies consistently outperform previous ones [1]. However, advanced methods may incur increased computational costs, necessitating a balance between costs and benefits [16].

Enhancing sample efficiency involves innovative frameworks, strategic human feedback utilization, and careful computational resource management. Techniques like RLHF and insights from recent studies, including the effective horizon concept, can lead to RL systems that adapt effectively to varying conditions and complexities in real-world applications [10, 36, 118, 13, 110].

8.3 Balancing Exploration and Exploitation

Balancing exploration and exploitation is fundamental in RL, requiring an optimal trade-off between exploring new actions and exploiting known actions to maximize rewards. This dilemma is pronounced in complex environments like robotics or gaming. Recent advancements, including intrinsically-motivated RL and adaptive exploration techniques such as axPPO, enhance exploration strategies by dynamically adjusting incentives based on agent performance or intrinsic motivations [59, 38, 119, 36].

Dynamic mechanisms within policy optimization frameworks, such as Pb-PPO's dynamic clipping strategy, enhance performance and stability by responding to task feedback [54]. Intrinsic motivation mechanisms enhance exploration capabilities, particularly in scenarios with limited or delayed external rewards, by quantifying state novelty and providing intrinsic rewards [26, 59, 96, 120].

Advanced methods like PPG contribute to balancing exploration and exploitation through decoupling policy and value function training, enabling precise optimization strategies. This dual approach improves sample efficiency and boosts learning performance, as demonstrated by recent advancements in RL algorithms like PPO and diffusion model integration [59, 36, 118, 71, 66].

Dynamic and adaptive techniques are crucial for achieving a successful balance between exploration and exploitation in RL, as evidenced by advancements in algorithms like SQIRL, which separate exploration from learning to enhance performance in stochastic environments. These techniques improve learning efficiency and facilitate RL applications in complex scenarios, including training large language models through methods like RLHF [118, 13, 10, 36]. By leveraging these innovative strategies, researchers can enhance the robustness and effectiveness of RL systems, enabling them to navigate complex environments with greater agility and precision.

8.4 Leveraging Human Feedback

Leveraging human feedback in RL is vital for enhancing algorithmic efficiency by aligning AI systems with human values and expectations. Integrating human feedback addresses challenges such as aligning AI outputs with user preferences and managing distributional gaps in off-policy training. Techniques like SPR reduce the memory footprint of policy optimization, aligning computational costs with standard supervised fine-tuning, thereby improving policy update efficiency [82].

Incorporating human feedback facilitates adaptive learning processes without relying on specific models, ensuring computational tractability and optimality. This flexibility enhances RL systems' adaptability and efficiency. For instance, PPO-RPE maintains a symmetric density ratio, aiding accurate error scale estimation and adaptive threshold design, refining the learning process [121].

Moreover, human feedback enhances the robustness of RL systems. Incorporating uncertainty into the learning process leads to more robust policy updates and improved alignment with human preferences [122]. This approach equips RL systems to handle real-world complexities, ultimately improving performance and reliability.

Human feedback is also pivotal in improving RL algorithms' performance in high-dimensional state spaces, evidenced by deep learning applications [8]. By dynamically adjusting intrinsic rewards to enhance extrinsic learning, methods like LIRPG demonstrate improved overall agent performance, highlighting human feedback's potential in refining RL frameworks [120].

The integration of human feedback can significantly reduce the amount required without compromising model performance, as seen in the APPO approach [92]. This reduction is crucial for scaling RL systems and ensuring their practical applicability across various domains.

Leveraging human feedback in RL is essential for enhancing algorithmic efficiency and ensuring AI models produce high-quality outputs aligned with human values and expectations. This approach, known as Reinforcement Learning from Human Feedback (RLHF), addresses undesirable behaviors in language models, such as generating false or irrelevant content, while incorporating fine-grained feedback for detailed insights into specific errors. Techniques like Parameter Efficient Reinforcement Learning from Human Feedback (PE-RLHF) reduce computational costs while maintaining performance, optimizing the training process. Embracing pluralism in RLHF can enhance AI systems' responsiveness to diverse human needs, leading to improved alignment and functionality across applications, from text generation to robotic navigation [25, 32, 27]. By addressing traditional RL limitations and incorporating diverse feedback mechanisms, researchers can develop more robust and adaptable AI solutions that closely align with human values and expectations.

8.5 Reducing Training Complexity

Reducing training complexity in RL is critical for enhancing scalability and applicability across diverse environments. The Weighted Policy Optimization (WPO) approach combines off-policy learning's efficiency with on-policy learning's performance benefits, optimizing the balance between exploration and exploitation to reduce computational burden [123]. The adaptive clipping approach in PPO controls the scale of policy updates based on state significance, improving learning reliability by ensuring updates are appropriately scaled [79]. Exploration-driven policy optimization techniques in RLHF further minimize required human feedback, leveraging preference feedback's structural advantages over numerical rewards to simplify learning [124].

Normality-guided distributional reinforcement learning enhances training times and reduces model complexity compared to ensemble-based methods, focusing on the most informative environmental aspects [125]. Model-free hierarchical learning representations simplify the learning process and enhance scalability by eliminating the need for a model of the environment, reducing computational demands [126].

Future research should refine existing methods, such as the logarithmic barrier method, and explore alternative distance measures to enhance performance across a broader range of environments [64]. Optimizing the SPR pipeline for larger LLMs and exploring enhancements to improve time and memory efficiency are crucial for reducing training complexity in RL models [90].

The discussed strategies underscore the potential to simplify training processes in RL, facilitating more efficient and scalable learning across various applications. A systematic review of RL-enhanced large language models (LLMs) highlights the complexities involved in implementing techniques like RLHF and DPO. While effective in improving model performance, these methods reveal challenges related to algorithmic complexity and resource allocation. Findings suggest that optimizing data diversity and volume can enhance reward model performance, indicating targeted strategies could mitigate inefficiencies observed in RLHF scaling. This understanding aids researchers in navigating current limitations and paves the way for future advancements in RL applications [10, 106]. By leveraging innovative methodologies and refining existing techniques, researchers can develop RL systems that are robust and adaptable to real-world challenges.

9 Conclusion

The investigation into Reinforcement Learning (RL) techniques underscores their transformative capacity across diverse fields, highlighting the necessity for innovative approaches to enhance the discipline's scalability and applicability. Central to these advancements are improvements in sample efficiency and the optimization of computational resources. The integration of human feedback, especially through Preference-based Reinforcement Learning (PbRL), offers a compelling alternative to traditional reward engineering by advocating for improved feedback mechanisms. Advanced policy optimization methods, including Maximum Entropy On-Policy Actor-Critic (EAPO) and Natural Policy Gradient (NPG), have significantly improved policy optimization performance and convergence rates.

Addressing gaps in learning objectives through entropy-regularized token-level policy optimization has led to substantial performance gains in interactive decision-making tasks. The amalgamation of large-scale language models (LLMs) with RL is crucial for fostering richer interactions between AI systems and human users, thereby ensuring AI outputs align with human preferences. Notably, the bilevel-LLM RLEF framework has shown significant progress in treatment performance and question-answering abilities.

Future research should focus on optimizing gradient estimators and examining the effects of various mixture model configurations on learning efficiency. Developing a unified linear programming framework could enhance sample efficiency and robustness in offline reward learning. Additionally, refining POME to better address sparse rewards, possibly through curiosity-driven exploration methods, represents a promising avenue for future exploration.

Experiments reveal that the effective horizon is a critical predictor of RL performance, providing a valuable framework for understanding the complexities of Markov Decision Processes (MDPs) in RL. The RLHC algorithm's superior performance compared to the PPO benchmark in competitive tasks highlights the advantages of hierarchical learning with multiple critics. Furthermore, the RISE method has demonstrated significant improvements in exploration efficiency and policy performance, affirming its superiority over existing approaches.

The current state of RL research presents a dynamic field with abundant opportunities for innovation and enhancement. By addressing existing challenges and embracing novel methodologies, researchers can develop more robust and effective RL systems that closely align with human values and expectations. Continued exploration of RL techniques promises to drive substantial advancements in artificial intelligence, ultimately leading to more sophisticated and human-centered AI solutions.

AI-generated, for reference only.

References

[1] Jia Hu, Xuerun Yan, Tian Xu, and Haoran Wang. Automated driving with evolution capability: A reinforcement learning method with monotonic performance enhancement, 2024.

[2] Thibaut Théate and Damien Ernst. Risk-sensitive policy with distributional reinforcement learning, 2022.

[3] William Koch, Renato Mancuso, Richard West, and Azer Bestavros. Reinforcement learning for uav attitude control, 2018.

[4] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization, 2023.

[5] Gabriel Mukobi, Peter Chatain, Su Fong, Robert Windesheim, Gitta Kutyniok, Kush Bhatia, and Silas Alberti. Superhf: Supervised iterative learning from human feedback, 2023.

[6] Mark Gluzman. Processing network controls via deep reinforcement learning, 2022.

[7] Zehong Cao and Chin-Teng Lin. Reinforcement learning from hierarchical critics, 2020.

[8] Luca Renna. Deep reinforcement learning for 2d physics-based object manipulation in clutter, 2023.

[9] Burcu Küçükoğlu, Walraaf Borkent, Bodo Rueckauer, Nasir Ahmad, Umut Güçlü, and Marcel van Gerven. Efficient deep reinforcement learning with predictive processing proximal policy optimization, 2024.

[10] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey, 2025.

[11] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment, 2024.

[12] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms, 2024.

[13] Hao Sun. Reinforcement learning in the era of llms: What is essential? what is needed? an rl perspective on rlhf, prompting, and beyond, 2023.

[14] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Accurate credit assignment in rl for llm mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

[15] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

[16] Cristian Bodnar, Ben Day, and Pietro Lió. Proximal distilled evolutionary reinforcement learning, 2020.

[17] Chang Yang, Ruiyu Wang, Xinrun Wang, and Zhen Wang. A game-theoretic perspective of generalization in reinforcement learning, 2022.

[18] Chenliang Li, Siliang Zeng, Zeyi Liao, Jiaxiang Li, Dongyeop Kang, Alfredo Garcia, and Mingyi Hong. Joint demonstration and preference learning improves policy alignment with human feedback, 2024.

[19] Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment, 2023.

AI-generated, for reference only.

[20] Sanghyun Son, Laura Yu Zheng, Ryan Sullivan, Yi-Ling Qiao, and Ming C. Lin. Gradient informed proximal policy optimization, 2023.

[21] Zifeng Zhuang, Kun Lei, Jinxin Liu, Donglin Wang, and Yilang Guo. Behavior proximal policy optimization, 2023.

[22] Weiqin Chen, James Onyejizu, Long Vu, Lan Hoang, Dharmashankar Subramanian, Koushik Kar, Sandipan Mishra, and Santiago Paternain. Adaptive primal-dual method for safe reinforcement learning, 2024.

[23] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf, 2024.

[24] Geetansh Kalra, Divye Singh, and Justin Jose. Rlinspect: An interactive visual approach to assess reinforcement learning algorithm, 2024.

[25] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training, 2023.

[26] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback, 2023.

[27] Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement learning from human feedback: Whose culture, whose values, whose perspectives?, 2025.

[28] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Improving reinforcement learning from human feedback using contrastive rewards, 2024.

[29] Zaifan Jiang, Xing Huang, and Chao Wei. Preference as reward, maximum preference optimization with importance sampling, 2024.

[30] Tianyi Qiu, Fanzhi Zeng, Jiaming Ji, Dong Yan, Kaile Wang, Jiayi Zhou, Yang Han, Josef Dai, Xuehai Pan, and Yaodong Yang. Reward generalization in rlhf: A topological perspective, 2024.

[31] W. Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi. Models of human preference for learning reward functions, 2023.

[32] Hakim Sidahmed, Samrat Phatale, Alex Hutcheson, Zhuonan Lin, Zhang Chen, Zac Yu, Jarvis Jin, Simral Chaudhary, Roman Komarytsia, Christiane Ahlheim, Yonghao Zhu, Bowen Li, Saravanan Ganesh, Bill Byrne, Jessica Hoffmann, Hassan Mansoor, Wei Li, Abhinav Rastogi, and Lucas Dixon. Parameter efficient reinforcement learning from human feedback, 2024.

[33] Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *arXiv preprint arXiv:2412.18279*, 2024.

[34] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Secrets of rlhf in large language models part ii: Reward modeling, 2024.

[35] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Robust rlhf with noisy rewards.

[36] Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards, 2020.

[37] Robust policy optimization in deep reinforcement learning.

[38] Andrei Lixandru. Proximal policy optimization with adaptive exploration, 2024.

[39] Karl Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient, 2020.

AI-generated, for reference only.

[40] Chris Lu, Jakub Grudzien Kuba, Alistair Letcher, Luke Metz, Christian Schroeder de Witt, and Jakob Foerster. Discovered policy optimisation, 2022.

[41] Shuang Feng and Grace Feng. An extremely data-efficient and generative llm-based reinforcement learning agent for recommenders, 2024.

[42] Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes, 2023.

[43] Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I. Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in rlhf, 2024.

[44] Yannick Metz, David Lindner, Raphaël Baur, and Mennatallah El-Assady. Mapping out the space of human feedback for reinforcement learning: A conceptual framework, 2025.

[45] Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanović, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences, 2024.

[46] Jiayi Zhou, Jiaming Ji, Juntao Dai, and Yaodong Yang. Sequence to sequence reward modeling: Improving rlhf by language feedback, 2024.

[47] Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles, 2023.

[48] Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

[49] Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. A theoretical framework for partially observed reward-states in rlhf, 2024.

[50] Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration, 2023.

[51] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization, 2024.

[52] Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. Rlhf-blender: A configurable interactive interface for learning from diverse human feedback, 2023.

[53] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms, 2024.

[54] Ziqi Zhang, Jingzehua Xu, Zifeng Zhuang, Hongyin Zhang, Jinxin Liu, Donglin wang, and Shuai Zhang. A dynamical clipping approach with task feedback for proximal policy optimization, 2024.

[55] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[56] Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization, 2019.

[57] Jayesh Singla, Ananye Agarwal, and Deepak Pathak. Sapg: Split and aggregate policy gradients, 2024.

[58] Ppo-ue: Proximal policy optimization via uncertainty-aware exploration.

[59] Mingqi Yuan. Intrinsically-motivated reinforcement learning: A brief introduction, 2022.

[60] Yaozhong Gan, Renye Yan, Zhe Wu, and Junliang Xing. Reflective policy optimization, 2024.

[61] Zhenyu Zhang, Xiangfeng Luo, Tong Liu, Shaorong Xie, Jianshu Wang, Wei Wang, Yang Li, and Yan Peng. Proximal policy optimization with mixed distributed training, 2019.

[62] Qining Zhang and Lei Ying. Zeroth-order policy gradient for reinforcement learning from human feedback without reward inference, 2024.

[63] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J. Zico Kolter, Zachary C. Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients, 2021.

[64] Cheng Zeng and Hongming Zhang. A logarithmic barrier method for proximal policy optimization, 2018.

[65] Zhengpeng Xie, Changdong Yu, and Weizheng Qiao. Dropout strategy in reinforcement learning: Limiting the surrogate objective variance in policy optimization methods, 2023.

[66] Junwei Zhang, Zhenghao Zhang, Shuai Han, and Shuai Lü. Proximal policy optimization via enhanced exploration efficiency, 2020.

[67] Wangshu Zhu and Andre Rosendo. Proximal policy optimization smoothed algorithm, 2020.

[68] Ju-Seung Byun, Byungmoon Kim, and Huamin Wang. Proximal policy gradient: Ppo with policy gradient, 2020.

[69] Sharan Vaswani, Amirreza Kazemi, Reza Babanezhad, and Nicolas Le Roux. Decision-aware actor-critic with function approximation and theoretical guarantees, 2023.

[70] Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization, 2025.

[71] Gao Tianci, Dmitriev D. Dmitry, Konstantin A. Neusypin, Yang Bo, and Rao Shengren. Enhancing sample efficiency and exploration in reinforcement learning through the integration of diffusion models and proximal policy optimization, 2025.

[72] Jean Seong Bjorn Choe and Jong-Kook Kim. Maximum entropy on-policy actor-critic via entropy advantage estimation, 2024.

[73] Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf, 2025.

[74] Dibya Ghosh, Marlos C. Machado, and Nicolas Le Roux. An operator view of policy gradient methods, 2020.

[75] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on ppo and trpo, 2020.

[76] Motoki Omura, Yasuhiro Fujita, and Toshiki Kataoka. Entropy controllable direct preference optimization, 2024.

[77] Jared Markowitz and Edward W. Staley. Clipped-objective policy gradients for pessimistic policy optimization, 2023.

[78] Feiyang Pan, Qingpeng Cai, An-Xiang Zeng, Chun-Xiang Pan, Qing Da, Hualin He, Qing He, and Pingzhong Tang. Policy optimization with model-based explorations, 2018.

[79] Gang Chen, Yiming Peng, and Mengjie Zhang. An adaptive clipping approach for proximal policy optimization, 2018.

[80] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study, 2024.

[81] Miao Fan, Chen Hu, and Shuchang Zhou. Proximal policy optimization actual combat: Manipulating output tokenizer length, 2023.

[82] Xinnan Zhang, Siliang Zeng, Jiaxiang Li, Kaixiang Lin, and Mingyi Hong. Policy optimization can be memory-efficient: Llm alignment through successive policy re-weighting (spr).

[83] Hari Prasad, Chinnu Jacob, and Imthias Ahamed T. P. Appraisal-guided proximal policy optimization: Modeling psychological disorders in dynamic grid world, 2024.

[84] Yi-Lin Tuan, Jinzhi Zhang, Yujia Li, and Hung yi Lee. Proximal policy optimization and its dynamic version for sequence generation, 2018.

[85] Hanyang Zhao, Genta Indra Winata, Anirban Das, Shi-Xiong Zhang, David D Yao, Wenpin Tang, and Sambit Sahu. Rainbowpo: A unified framework for combining improvements in preference optimization. *arXiv preprint arXiv:2410.04203*, 2024.

[86] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of rlhf in large language models part i: Ppo, 2023.

[87] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf, 2024.

[88] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[89] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback, 2024.

[90] Xinnan Zhang, Siliang Zeng, Jiaxiang Li, Kaixiang Lin, and Mingyi Hong. Llm alignment through successive policy re-weighting (spr). In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

[91] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation, 2024.

[92] Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries, 2025.

[93] Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback, 2024.

[94] Shiming Xie, Hong Chen, Fred Yu, Zeye Sun, Xiuyu Wu, and Yingfan Hu. Minor dpo reject penalty to increase training robustness, 2024.

[95] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024.

[96] Yong Lin, Skyler Seto, Maartje ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. On the limited generalization capability of the implicit reward model induced by direct preference optimization, 2024.

[97] Yongtao Wu, Luca Viano, Yihang Chen, Zhenyu Zhu, Quanquan Gu, and Volkan Cevher. Multi-step preference optimization via two-player markov games. In *Language Gamification-NeurIPS 2024 Workshop*.

[98] Moritz Stephan, Alexander Khazatsky, Eric Mitchell, Annie S Chen, Sheryl Hsu, Archit Sharma, and Chelsea Finn. Rlvf: Learning from verbal feedback without overgeneralization, 2024.

AI-generated, for reference only.

[99] Alexander G. Padula and Dennis J. N. J. Soemers. Exploring rl-based llm training for formal language tasks with programmed rewards, 2024.

[100] Nathan Lambert and Roberto Calandra. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback, 2024.

[101] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024.

[102] Alexey Gorbatovski and Sergey Kovalchuk. Reinforcement learning for question answering in programming domain using public community scoring as a human feedback, 2024.

[103] Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in reinforcement learning from human feedback, 2024.

[104] Muning Wen, Junwei Liao, Cheng Deng, Jun Wang, Weinan Zhang, and Ying Wen. Entropy-regularized token-level policy optimization for language agent reinforcement, 2024.

[105] Zhenyu Hou, DU Pengfan, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. Does rlhf scale? exploring the effects of data, model, and method.

[106] Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. Does rlhf scale? exploring the impacts from data, model, and method, 2024.

[107] Michael Santacroce, Yadong Lu, Han Yu, Yuanzhi Li, and Yelong Shen. Efficient rlhf: Reducing the memory usage of ppo, 2023.

[108] Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning, 2024.

[109] Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The n+ implementation details of rlhf with ppo: A case study on tl;dr summarization, 2024.

[110] Cassidy Laidlaw, Stuart Russell, and Anca Dragan. Bridging rl theory and practice with the effective horizon, 2024.

[111] Ryan Sullivan, Akarsh Kumar, Shengyi Huang, John P. Dickerson, and Joseph Suarez. Reward scale robustness for proximal policy optimization via dreamerv3 tricks, 2023.

[112] Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards, 2024.

[113] Eugene Bykovets, Yannick Metz, Mennatallah El-Assady, Daniel A. Keim, and Joachim M. Buhmann. How to enable uncertainty estimation in proximal policy optimization, 2022.

[114] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[115] Atish Dixit and Ahmed Elsheikh. A multilevel reinforcement learning framework for pde-based control, 2022.

[116] Ching Pui Wan, Tung Li, and Jason Min Wang. Rlor: A flexible framework of deep reinforcement learning for operation research, 2023.

[117] Xin Chen, Sam Toyer, and Florian Shkurti. Exploring and addressing reward confusion in offline preference learning, 2024.

[118] Cassidy Laidlaw, Banghua Zhu, Stuart Russell, and Anca Dragan. The effective horizon explains deep rl performance in stochastic environments, 2024.

[119] Mohamed-Amine Chadi and Hajar Mousannif. Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization, 2023.

[120] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods, 2018.

[121] Taisuke Kobayashi. Proximal policy optimization with adaptive threshold for symmetric relative density ratio, 2022.

[122] Sam Houliston, Alizée Pace, Alexander Immer, and Gunnar Rätsch. Uncertainty-penalized direct preference optimization. *arXiv preprint arXiv:2410.20187*, 2024.

[123] Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization, 2024.

[124] Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Exploration-driven policy optimization in rlhf: Theoretical insights on efficient data utilization, 2024.

[125] Ju-Seung Byun and Andrew Perrault. Normality-guided distributional reinforcement learning for continuous control, 2024.

[126] Jacob Rafati and David C. Noelle. Learning representations in model-free hierarchical reinforcement learning, 2019.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.