# Introduction to Reinforcement Learning from Human Feedback: A Review of Current Developments

Satyadhar Joshi *

Posted Date: 17 March 2025

doi: 10.20944/preprints202503.1159.v1

Keywords: Reinforcement Learning from Human Feedback (RLHF); Large Language Models (LLMs); Reward Models; RLAIF; Safe RLHF; Direct Preference Optimization (DPO); Online RLHF

 doi:10.20944/preprints202503.1159.v1



*Article*

# Introduction to Reinforcement Learning from Human Feedback: A Review of Current Developments

**Satyadhar Joshi**

Independent Researcher, BoFA, Jersey City, NJ, USA; satyadhar.joshi@gmail.com

**Abstract:** Reinforcement Learning from Human Feedback (RLHF) has emerged as a pivotal technique for aligning large language models (LLMs) with human preferences. This paper provides a comprehensive overview of RLHF, examining its methodologies, challenges, and recent advancements. We analyze various approaches, including reward modeling, preference optimization, and the integration of AI feedback. This paper also provides theoretical foundations, practical implementations, and challenges. We explore the evolution of RLHF, its comparison with Reinforcement Learning from AI Feedback (RLAIF), and the role of reward models in optimizing LLMs. Additionally, we discuss recent advancements such as Safe RLHF, Direct Preference Optimization (DPO), and the integration of RLHF with online learning frameworks. The paper concludes with future directions and open problems in RLHF research. We also discuss the practical aspects of implementing RLHF, covering workflow from data collection to online training.

**Keywords:** Reinforcement Learning from Human Feedback (RLHF); Large Language Models (LLMs); Reward Models; RLAIF; Safe RLHF; Direct Preference Optimization (DPO); Online RLHF

## 1. Introduction

Reinforcement Learning from Human Feedback (RLHF) has gained significant attention as a method for aligning large language models (LLMs) with human preferences [1]. RLHF leverages human-generated feedback to fine-tune models, ensuring that their outputs are more aligned with human values and intentions. This approach has been particularly effective in mitigating issues such as toxicity and hallucinations in LLMs [2]. The success of RLHF has led to the development of various extensions and improvements, such as Reinforcement Learning from AI Feedback (RLAIF) [3], which uses AI-generated feedback to reduce the reliance on human annotations. However, RLHF still faces several challenges, including the high cost of human feedback, the difficulty of scaling to complex tasks, and the potential for reward model misspecification [4]. This paper aims to provide a comprehensive review of RLHF, covering its theoretical foundations, practical implementations, and recent advancements. We also discuss the limitations of current approaches and propose future directions for research in this area.

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text. However, aligning these models with human values and preferences remains a significant challenge. Reinforcement Learning from Human Feedback (RLHF) has become a key methodology in this area, enabling LLMs to better understand and respond to user intentions [5,6]. This paper aims to provide a comprehensive survey of RLHF, exploring its fundamental concepts, recent advancements, and practical considerations. RLHF combines traditional reinforcement learning with human-generated feedback to train a reward model that reflects human preferences [1,7–9]. This reward model is then used to optimize the LLM's policy. The process typically involves:

- Collecting human preference data
- Training a reward model
- Fine-tuning the LLM using reinforcement learning

Recent research has also explored AI feedback and alternative optimization methods [3,10,11].

 doi:10.20944/preprints202503.1159.v1



## 2. Recent Advancements in RLHF and GenAI

Recent advancements in RLHF have focused on improving the scalability and efficiency of the training process. For example, Direct Preference Optimization (DPO) has been proposed as an alternative to traditional RLHF, eliminating the need for a separate reward model by directly optimizing the policy using human preferences [10]. This approach has been shown to achieve comparable performance to RLHF while reducing computational costs [10].

Another significant advancement is the integration of RLHF with online learning frameworks. Online iterative RLHF, which involves continuous feedback collection and model updates, has been shown to outperform offline RLHF in terms of both performance and scalability [12]. This approach has been successfully implemented in large-scale LLM training pipelines, achieving state-of-the-art performance on benchmarks such as AlpacaEval-2 and MT-Bench [12].

Several studies have explored the integration of Generative AI into financial risk modeling. Joshi [13] introduced an enhanced Vasicek framework utilizing agentic generative AI for improved risk assessment. In another work, Joshi [14] conducted a comprehensive review of AI agent frameworks, highlighting their applications and challenges in financial stability.

Joshi's book [15] provides an in-depth analysis of AI-driven risk management techniques. Further, structured finance risk models such as Leland-Toft and Box-Cox were enhanced using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) in [16]. This study demonstrated the applicability of GenAI in refining financial risk assessment models.

Additionally, the robustness of the U.S. financial and regulatory system was examined through the implementation of GenAI techniques in [17]. A broader perspective on GenAI models for financial risk management was provided in [18], reviewing different approaches and their effectiveness.

Joshi [19] discussed the synergy between Generative AI and big data in financial risk modeling, emphasizing recent developments. Lastly, the role of prompt engineering in enhancing financial market integrity and risk management was analyzed in [20], showing how AI-driven methodologies improve decision-making processes.

These works collectively contribute to the growing body of research on AI-driven financial risk management, demonstrating the potential of Generative AI in improving predictive accuracy and stability in financial markets.

## 3. Theoretical Foundations and Methodology

RLHF builds on the principles of reinforcement learning, where an agent learns to maximize a reward signal through interaction with an environment. In the context of LLMs, the reward signal is derived from human feedback, which is used to guide the model's behavior [21]. The theoretical underpinnings of RLHF are rooted in preference-based learning, where human preferences are used to define the reward function [22].

Recent work has explored the theoretical differences between RLHF and standard reinforcement learning, demonstrating that RLHF can be reduced to robust reward-based RL under certain conditions [22]. This theoretical insight has practical implications for the design of RLHF algorithms, as it allows for the use of existing RL techniques with minimal modifications [22].

RLHF involves several key components, including reward modeling, policy optimization, and human feedback collection. The process typically begins with the collection of human preference data, which is used to train a reward model. This reward model is then used to guide the optimization of the LLM through reinforcement learning algorithms such as Proximal Policy Optimization (PPO) [23].

One of the challenges in RLHF is the design of the reward model. Traditional approaches rely on pairwise comparisons, where humans are asked to choose between two model outputs. However, this method can be labor-intensive and costly, especially for complex tasks [24]. To address this, recent work has proposed the use of groupwise comparisons and interactive feedback mechanisms to improve the efficiency of human feedback collection [24].

 doi:10.20944/preprints202503.1159.v1



Another important aspect of RLHF is the alignment of the reward model with human values. This requires careful consideration of the trade-offs between different objectives, such as helpfulness and harmlessness. Safe RLHF has been proposed as a solution to this problem, decoupling human preferences for helpfulness and harmlessness and training separate reward and cost models [25].

### *3.1. Practical Considerations*

Implementing RLHF involves several practical steps, from data collection to online training [12,26,27]. The workflow typically includes:

- Data collection and annotation
- Reward model training
- Online RLHF training

Open-source tools and platforms facilitate the implementation of RLHF [28].

## 4. Recent Advancements and Challenges

Recent advancements in Reinforcement Learning from Human Feedback (RLHF) have focused on improving scalability, efficiency, and alignment of large language models (LLMs). This section highlights key developments from the last 12 months, drawing on the latest research published in 2025. Researchers have explored various aspects of RLHF, including reward modeling [29,30], optimization techniques [23], and safety considerations [25]. Studies have also critically analyzed the complexities and limitations of RLHF [2,4,22].

### *4.1. Reward Modeling*

The accuracy of the reward model is crucial for effective RLHF. Evaluation of reward models has been explored [30]. Techniques for constructing and refining reward models have been proposed [29].

### *4.2. Optimization Techniques*

Direct Preference Optimization (DPO) and Proximal Policy Optimization (PPO) are commonly used optimization techniques [10,23]. These methods aim to efficiently align the LLM with the learned reward model.

### *4.3. Safety and Alignment*

Ensuring the safety and ethical alignment of LLMs is a critical aspect of RLHF. Approaches like SAFE RLHF aim to address these concerns [25].

### *4.4. Methodological Advances*

Reinforcement Learning from Human Feedback (RLHF) has become critical for aligning large language models (LLMs) with human values [1]. Core technical approaches include:

- Proximal Policy Optimization (PPO) implementations for RLHF [23]
- AI Feedback (RLAIF) achieving parity with human feedback [3]
- Direct Preference Optimization (DPO) bypassing explicit reward modeling [10]
- SAFE RLHF decoupling helpfulness/harmlessness objectives [25]

### *4.5. Theoretical Insights*

Recent analyses reveal fundamental challenges:

- Theoretical equivalence to standard RL problems [22]
- Reward model misspecification risks [2]
- Preference modeling as game-theoretic equilibrium [31]

### *4.6. Applications & Impact*

Practical implementations demonstrate RLHF's versatility:

 doi:10.20944/preprints202503.1159.v1



- Grammar correction systems [32]
- ChatGPT's human-computer interaction [33]
- LLM safety alignment [34]

### *4.7. Open Challenges*

Current research identifies key limitations:

- Sparse feedback constraints [21]
- Scalability of human preference collection [35]
- Generalization beyond training distributions [2]

### *4.8. Targeted Human Feedback for LLM Alignment*

One of the most significant advancements in 2025 is the introduction of **RLTHF** (Targeted Human Feedback for LLM Alignment) [36]. RLTHF addresses the high cost of human annotations by combining LLM-based initial alignment with selective human corrections. This hybrid approach identifies hard-to-annotate samples using a reward model's reward distribution and iteratively enhances alignment with minimal human effort. Evaluations on the HH-RLHF and TL;DR datasets demonstrate that RLTHF achieves full-human annotation-level alignment with only 6-7% of the human annotation effort. Furthermore, models trained on RLTHF-curated datasets outperform those trained on fully human-annotated datasets, underscoring the effectiveness of strategic data curation. The advancements in RLHF in 2025 have significantly improved the scalability, efficiency, and alignment of LLMs. Techniques such as RLTHF, online iterative RLHF, and enhanced reward modeling have addressed key challenges, including the high cost of human feedback and the generalization of reward models. As RLHF continues to evolve, these advancements will play a crucial role in the development of safe and effective AI systems.

### *4.9. Online Iterative RLHF*

Another major development in 2025 is the widespread adoption of **Online Iterative RLHF**. Unlike traditional offline RLHF, online iterative RLHF involves continuous feedback collection and model updates, enabling dynamic adaptation to evolving human preferences. This approach has been successfully implemented in large-scale LLM training pipelines, achieving state-of-the-art performance on benchmarks such as AlpacaEval-2, Arena-Hard, and MT-Bench. The integration of proxy preference models, which approximate human feedback using open-source datasets, has further reduced the reliance on costly human annotations.

### *4.10. Comprehensive Surveys on LLM Alignment Techniques*

In 2025, several comprehensive surveys have been published, providing a detailed overview of LLM alignment techniques, including RLHF, RLAIF, PPO, and DPO [37]. These surveys categorize and explain various alignment methods, helping researchers and practitioners gain a thorough understanding of the field. They also highlight the limitations of current approaches, such as the reliance on human feedback and the challenges of reward model generalization, while proposing future directions for research.

### *4.11. Advancements in Reward Modeling*

Reward modeling has seen significant improvements in 2025, with the introduction of contrastive learning and meta-learning techniques to enhance the generalization capabilities of reward models [29]. These methods enable reward models to better distinguish between chosen and rejected responses, even for out-of-distribution samples. Additionally, meta-learning has been employed to maintain the reward model's ability to differentiate subtle differences during iterative RLHF optimization, further improving alignment with human preferences.

 doi:10.20944/preprints202503.1159.v1



### *4.12. Integration with Cloud Platforms*

The integration of RLHF with cloud platforms has also gained traction in 2025. For example, Google Cloud has introduced tools for tuning models like PaLM 2 and Llama 2 using RLHF [27]. These tools enable researchers and developers to leverage scalable infrastructure for RLHF training, reducing the computational burden and accelerating the development of aligned LLMs.

## 5. Methodologies

### *5.1. Reward Modeling and Training*

The core RLHF methodology involves three key components:

- **Preference Collection:** Human annotators rank model outputs to create training data [1]
- **Reward Model (RM) Training:** Neural networks learn to predict human preferences [23]
- **Policy Optimization:** Proximal Policy Optimization (PPO) fine-tunes LLMs using RM predictions [31]

### *5.2. Policy Optimization Techniques*

Recent advances in RLHF optimization include:

- **PPO Implementations:** Scalable methods for stable policy updates [23]
- **Direct Preference Optimization (DPO):** Bypasses explicit reward modeling through contrastive learning [10]
- **RLAIF Variants:** AI-generated feedback reduces human annotation costs [3]

### *5.3. Safety Mechanisms*

Critical safety components address alignment challenges:

- **SAFE RLHF:** Lagrangian optimization balances helpfulness/harmlessness [25]
- **Multi-Objective RL:** Decoupled reward modeling for different safety aspects [2]

### *5.4. Theoretical Frameworks*

Methodological foundations build on:

- **Game-Theoretic Analysis:** Modeling preferences as Nash equilibria [22]
- **RL Equivalence Proofs:** Formal reduction to standard RL problems [21]

### *5.5. Practical Implementation*

Deployment strategies include:

- **Iterative Refinement:** Continuous human-AI feedback loops [35]
- **Domain-Specific Adaptation:** Customization for applications like grammar correction [32]
- **Human Interaction Design:** ChatGPT-style conversational interfaces [33]

## 6. Applications of RLHF

RLHF has been applied to various applications, including grammar error correction [32] and improving the overall quality of LLM outputs. The impact of RLHF extends to various domains, including human-computer interaction [33] and data management [38].

RLHF has been applied to a wide range of tasks, including grammar error correction, summarization, and dialogue generation. For example, RLHF has been used to fine-tune models for non-native English speakers, significantly improving their ability to correct grammatical errors [32]. In the context of summarization, RLHF has been shown to produce summaries that are more aligned with human preferences compared to traditional supervised fine-tuning [3].

In addition to these applications, RLHF has also been used to improve the safety and alignment of LLMs. For example, Safe RLHF has been proposed as a method for ensuring that LLMs generate

 doi:10.20944/preprints202503.1159.v1



helpful and harmless responses [25]. This approach has been shown to outperform existing algorithms in terms of both performance and safety [25].

## 7. Applications of Reinforcement Learning from Human Feedback

### *7.1. Enhancing Language Model Capabilities*

RLHF has found widespread use in improving various aspects of language model performance.

- **Dialogue Generation:** RLHF is used to train dialogue agents that are more helpful and harmless, aligning them with human preferences [3,25]. This includes generating responses that are both informative and avoid harmful content.
- **Summarization:** RLAIF demonstrates comparable results to RLHF using AI-generated feedback for summarization tasks, reducing the need for expensive human annotation [3].
- **Grammar Error Correction:** Trinka AI utilizes RLHF to develop grammar error correction models specifically designed for non-native English speakers [32].
- **Improving Reasoning Capabilities:** RLHF enhances the reasoning capabilities of LLMs, facilitating their use as human-centric assistants [23].

### *7.2. Human-Computer Interaction*

RLHF plays a crucial role in shaping the interaction between humans and language models.

- **ChatGPT and User Experience:** [33] highlights how RLHF in ChatGPT improves conversational interfaces and the overall user experience, making interactions more natural and human-like.
- **Customer Service Applications:** RLHF is being applied to develop more effective and efficient customer service chatbots, improving customer satisfaction and reducing support costs [33].

### *7.3. Theoretical Application and Analysis*

- **Theoretical Examination of RLHF:** RLHF's applications have led to increased study of it as a concept, as seen by its theoretical examination and comparison against standard RL algorithms [22].

### *7.4. Customer Service and Financial Advice*

Drawing parallels from the demonstrated success of RLHF in improving customer service chatbots [33], one could envision LLMs trained with RLHF to provide more personalized and helpful financial advice to customers. The models could learn to understand individual financial situations, risk tolerance, and investment goals through human feedback, leading to more tailored recommendations.

## 8. Gap Analysis and Future Directions for Reinforcement Learning from Human Feedback (RLHF)

While Reinforcement Learning from Human Feedback (RLHF) has demonstrated remarkable success in aligning large language models (LLMs) with human preferences [1,3,33], significant gaps and challenges remain that warrant further investigation. This section identifies these gaps and suggests potential avenues for future research.

### *8.1. Challenges and Limitations*

Despite its successes, RLHF still faces several challenges. One of the main limitations is the reliance on human feedback, which can be expensive and time-consuming to collect. This has led to the development of alternative approaches such as RLAIF, which uses AI-generated feedback to reduce the need for human annotations [3]. However, RLAIF also has its limitations, including the potential for bias in the AI-generated feedback [3].

Another challenge is the generalization of reward models to out-of-distribution data. Reward models trained on a specific dataset may struggle to generalize to new tasks or domains, leading to

 doi:10.20944/preprints202503.1159.v1



suboptimal performance [29]. To address this, recent work has proposed the use of contrastive learning and meta-learning to improve the generalization capabilities of reward models [29].

Finally, the ethical implications of RLHF must be carefully considered. The use of human feedback to train AI systems raises concerns about privacy, consent, and the potential for misuse. Future research should focus on developing methods for ensuring the ethical use of RLHF, including the development of transparent and accountable feedback collection processes [4].

### *8.2. Limitations of Current Methodologies*

#### 8.2.1. Reward Model Misspecification

A critical challenge lies in the potential for reward model misspecification, where the learned reward function fails to accurately capture human preferences [2]. This can lead to unintended consequences, such as reward hacking or the generation of outputs that are superficially aligned with human feedback but lack genuine understanding or ethical considerations. More robust reward modeling techniques are needed to address this limitation. Works such as [23] can help address the model design, however, new methods in model evaluation are required to address the underlying limitations.

#### 8.2.2. Sparse and Biased Feedback

RLHF relies on human feedback, which can be sparse, noisy, and biased [2]. The cost and scalability of collecting high-quality human feedback remain significant barriers to wider adoption. The limitations of biased information are touched upon by [21] and [35].

### *8.3. Safety and Ethical Concerns*

#### 8.3.1. Safety Alignment

Ensuring the safety and ethical alignment of LLMs trained with RLHF remains a paramount concern [25]. While techniques like SAFE RLHF [25] offer promising approaches, further research is needed to develop more robust and reliable safety mechanisms that can prevent the generation of harmful, biased, or misleading content. The theoretical concepts behind alignment are discussed in [22] and must be coupled with the practical implications found in [35] to develop robust safety mechanisms.

#### 8.3.2. Scalability

Applying RLHF to customer service requires scalability to meet the requirements [33] of customers. More efficient methods should be developed to address this gap.

#### 8.3.3. Self-Improvement and AI Feedback

Exploring the potential of self-improvement through AI feedback (RLAIF) represents a promising avenue for future research [3]. Developing techniques that enable LLMs to learn from their own outputs and iteratively refine their behavior could significantly reduce the reliance on human feedback and improve the scalability of RLHF. Further exploration into the theoretical and practical implications could be explored using [10].

#### 8.3.4. Theoretical Understanding

Gaining a deeper theoretical understanding of RLHF is crucial for developing more effective and reliable algorithms [22,31]. Further research is needed to analyze the convergence properties of RLHF algorithms, understand the impact of different preference models, and develop theoretical guarantees for safety and performance. This includes closing the gaps identified by [34].

## 9. Analysis of Web Articles on Reinforcement Learning from Human Feedback

This section analyzes web articles from the provided bibliography, summarizing their content and assessing their value in understanding Reinforcement Learning from Human Feedback (RLHF).

 doi:10.20944/preprints202503.1159.v1



These articles offer accessible overviews and practical insights, complementing the more technical research papers.

*9.1. General Overviews*

9.1.1. Swimm.io: What Is Reinforcement Learning from Human Feedback (RLHF)? [1]

This article on Swimm.io provides a high-level introduction to RLHF, explaining its core concepts and how it combines traditional reinforcement learning with human-generated feedback. It serves as a valuable starting point for readers unfamiliar with the topic.

9.1.2. ResearchGate: (PDF) 7 Reinforcement Learning from Human Feedback (RLHF) [21]

This entry on ResearchGate links to a PDF document providing an overview of RLHF. ResearchGate is a collaborative place for scientists to share and discuss projects.

9.1.3. GitHub: RLHF/main.pdf at main · Peymankor/RLHF [35]

This resource, hosted on GitHub, offers a tutorial on RLHF, focusing on aligning AI behavior with human values through feedback-driven training. It likely contains practical code examples and implementation details.

*9.2. Specific Applications and Techniques*

9.2.1. Trinka AI: RLHF for Grammar Error Correction [32]

This technical blog post from Trinka AI describes their use of RLHF to develop a Grammar Error Correction (GEC) model specifically designed for non-native speakers of English. It offers insights into the practical application of RLHF in a specific domain.

*9.3. Limitations and Future Directions*

While these articles provide valuable overviews and practical examples, they generally lack the depth and rigor of peer-reviewed research papers. For a more critical analysis of the limitations of RLHF and potential future directions, readers should consult the research articles cited elsewhere in this review (e.g., [2]).

## 10. Architectural Considerations in Reinforcement Learning from Human Feedback Systems

This section explores the architectural design and implementation aspects of Reinforcement Learning from Human Feedback (RLHF) systems. It draws upon the provided literature to highlight key components and design choices that influence the performance, scalability, and safety of these systems.

*10.1. Core System Architecture*

A typical RLHF system comprises several key modules that interact to align LLMs with human preferences:

- **Large Language Model (LLM):** The foundation of the system, responsible for generating text outputs [1].
- **Human Feedback Mechanism:** A means for collecting human preferences, typically through ranking or rating model outputs [23].
- **Reward Model (RM):** A neural network trained to predict human preferences based on the collected feedback [23]. The importance of the RM is often discussed as an important aspect of design when implementing an RLHF system. The RM enables more stable design when implementing an RLHF system.
- **Policy Optimization Algorithm:** An RL algorithm, such as Proximal Policy Optimization (PPO), used to fine-tune the LLM based on the reward signal from the RM [23,31].

 doi:10.20944/preprints202503.1159.v1



### *10.2. Architectural Variants and Optimizations*

Researchers have explored various architectural modifications and optimizations to improve the efficiency and effectiveness of RLHF systems:

- **Direct Preference Optimization (DPO):** DPO bypasses explicit reward modeling by directly optimizing the LLM policy based on preference data [10].
- **RL from AI Feedback (RLAIF):** RLAIF replaces human feedback with feedback generated by another AI model, potentially reducing the cost and scalability challenges associated with human annotation [3]. This is not to say that the architectural benefits are without limitation, as discussed in [2].

### *10.3. Safety and Alignment Architectures*

Ensuring the safety and ethical alignment of RLHF systems requires careful architectural considerations:

- **SAFE RLHF:** This approach decouples helpfulness and harmlessness objectives, using Lagrangian optimization to balance these competing goals [25]. By explicitly modeling and controlling for safety constraints, SAFE RLHF aims to mitigate the risk of generating harmful or unethical content.

### *10.4. Integration Considerations*

- **Chatbot Integration:** With RLHF being discussed as a crucial element to the design of Chatbots [33], architectural design must be considered when implementing RLHF and chatbots together.

### *10.5. Scalability*

Architectural decisions for systems using RLHF need to account for scalability [35] to meet customer needs.

## 11. Architectural Considerations in RLHF

The architectural design of RLHF systems plays a pivotal role in their efficacy. This section delves into the key architectural components and considerations that influence the performance and scalability of RLHF.

### *11.1. Reward Model Architecture*

The reward model, a core component of RLHF, is responsible for predicting human preferences. Its architecture significantly impacts the accuracy and efficiency of the learning process. Recent studies have explored various architectures, including transformer-based models, to capture complex human preferences [29]. The choice of architecture must balance the need for expressive power with computational efficiency.

### *11.2. Policy Optimization Architecture*

The policy optimization architecture determines how the LLM's policy is updated based on the reward signal. Techniques like Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) necessitate specific architectural considerations. PPO, for example, often involves a separate actor and critic network [23]. DPO, on the other hand, reformulates the RLHF problem as a classification task, simplifying the architecture [10].

### *11.3. Feedback Collection Architecture*

The architecture for collecting human feedback is crucial for gathering high-quality preference data. This includes the design of user interfaces, data annotation platforms, and methods for ensuring data diversity and reliability. Interactive groupwise comparisons have been explored to increase feedback efficiency [24].

 doi:10.20944/preprints202503.1159.v1



*11.4. Scalability and Distributed Training*

As LLMs continue to grow in size, the scalability of RLHF becomes paramount. Distributed training architectures, including data and model parallelism, are essential for handling the computational demands of RLHF. Cloud-based platforms and distributed computing frameworks are increasingly used to scale RLHF training [27].

*11.5. Integration with LLM Architecture*

The integration of RLHF with the underlying LLM architecture is a critical consideration. This includes the design of interfaces for passing reward signals and updating model parameters. The architectural design must also account for the potential impact of RLHF on the LLM's overall performance and safety.

*11.6. RLAIF Architecture*

The introduction of Reinforcement Learning from AI Feedback (RLAIF) introduces new architectural designs where AI models are used to provide feedback. This requires an architecture that can effectively integrate and process AI-generated feedback alongside or in place of human input [3,11].

## 12. Quantitative Foundations for RLHF

This section analyzes the provided literature for the presence and use of quantitative mathematics and equations in the context of Reinforcement Learning from Human Feedback (RLHF). While some sources offer conceptual overviews, others delve into the mathematical formalisms underlying RLHF algorithms and methodologies. These equations are essential for understanding the underlying principles and algorithms used in RLHF.

*12.1. Theoretical Foundations*

12.1.1. Wang et al.: Is RLHF More Difficult than Standard RL? A Theoretical Perspective [22]

This paper uses quantitative analysis and equations to theorectically prove the effectiveness of preference-based RL.

*12.2. Core Methodologies*

12.2.1. Proximal Policy Optimization (PPO) [23,31]

The core paper by Zheng et al, and summarized by Lambert implements equations with PPO algorithms for optimizing the human-centric reward functions. The PPO algorithm itself is rooted in mathematical equations for policy updates, Kullback–Leibler (KL) divergence control, and value function estimation.

12.2.2. Reward Modeling

While not explicitly detailing the equations, many articles discuss building an accurate RM [23] as important to RLHF design.

*12.3. Limitations*

Though many architectural designs are theoretically sound [22] they have limitations as discussed by Chaudhari.

*12.4. Reward Model Formulation*

The reward model $r_\theta(x, y)$ is typically trained to predict the human preference between two outputs $y_1$ and $y_2$ given an input $x$. This is often formulated as a binary classification problem, where:

$$P(y_1 \succ y_2|x) \approx \sigma(r_\theta(x, y_1) - r_\theta(x, y_2)) \tag{1}$$

 doi:10.20944/preprints202503.1159.v1



where $\sigma$ is the sigmoid function, and $y_1 \succ y_2$ indicates that $y_1$ is preferred over $y_2$. The reward model parameters $\theta$ are trained to minimize the cross-entropy loss:

$$\begin{aligned}\mathcal{L}(\theta) = -\mathbb{E}_{(x,y_1,y_2,\text{label})}[&\text{label} \log P(y_1 \succ y_2|x) \\ &+ (1 - \text{label}) \log(1 - P(y_1 \succ y_2|x))]\end{aligned} \tag{2}$$

where label is 1 if $y_1 \succ y_2$ and 0 otherwise.

*12.5. Proximal Policy Optimization (PPO)*

PPO is a commonly used RL algorithm in RLHF. The objective is to maximize the expected reward while ensuring that the policy updates are not too large. The PPO objective function is:

$$\mathcal{L}^{\text{CLIP}}(\phi) = \mathbb{E}_t[\min(r_t(\phi)\hat{A}_t, \text{clip}(r_t(\phi), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \tag{3}$$

where $r_t(\phi) = \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{\text{old}}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the advantage estimate, $\epsilon$ is a clipping parameter, and $\pi_\phi$ is the policy with parameters $\phi$ [23].

*12.6. Direct Preference Optimization (DPO)*

DPO reformulates RLHF as a classification problem. The DPO objective is:

$$\mathcal{L}^{\text{DPO}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\left(\frac{\exp(\pi_\phi(y_w|x)/\beta)}{\sum_{y\in\{y_w,y_l\}}\exp(\pi_\phi(y|x)/\beta)}\right)\right] \tag{4}$$

where $y_w$ is the preferred (winner) output, $y_l$ is the dispreferred (loser) output, $\pi_\phi$ is the policy, and $\beta$ is a temperature parameter [10].

*12.7. Mathematical Challenges and Considerations*

The mathematical formulations in RLHF also highlight several challenges, such as the non-convex nature of the optimization problems and the potential for reward hacking. Theoretical perspectives on the difficulty of RLHF compared to standard RL have been explored [22].

## 13. Quantitative Formulations of RLHF

Reinforcement Learning from Human Feedback (RLHF) relies on mathematical formulations to model human preferences and optimize large language models (LLMs). This section provides an overview of the key mathematical concepts and equations used in RLHF, as described in recent literature.

The mathematical formulations of RLHF provide a rigorous framework for aligning LLMs with human preferences. From reward modeling and policy optimization to safe RLHF and online iterative RLHF, these equations capture the key principles and challenges of RLHF. Future research will continue to refine these formulations to improve the scalability, efficiency, and safety of RLHF.

*13.1. Reward Modeling*

The core component of RLHF is the reward model, which is trained to predict human preferences. Given a pair of model outputs $(y_1, y_2)$, the reward model $R_\theta$ assigns a scalar reward $R_\theta(y)$ to each output. The probability that $y_1$ is preferred over $y_2$ is modeled using the Bradley-Terry model:

$$P(y_1 \succ y_2) = \frac{\exp(R_\theta(y_1))}{\exp(R_\theta(y_1)) + \exp(R_\theta(y_2))} \tag{5}$$

where $R_\theta(y)$ is the reward for output $y$, and $\theta$ represents the parameters of the reward model [2].

 doi:10.20944/preprints202503.1159.v1



*13.2. Policy Optimization*

Once the reward model is trained, the policy $\pi_\phi$ is optimized using reinforcement learning algorithms such as Proximal Policy Optimization (PPO). The objective is to maximize the expected reward:

$$J(\phi) = \mathbb{E}_{y\sim\pi_\phi}[R_\theta(y)] - \beta\,\mathrm{KL}(\pi_\phi||\pi_{\text{ref}}) \tag{6}$$

where $\mathrm{KL}(\pi_\phi||\pi_{\text{ref}})$ is the Kullback-Leibler (KL) divergence between the current policy $\pi_\phi$ and a reference policy $\pi_{\text{ref}}$, and $\beta$ is a regularization parameter that controls the deviation from the reference policy [23].

*13.3. Direct Preference Optimization (DPO)*

Direct Preference Optimization (DPO) is an alternative to traditional RLHF that eliminates the need for a separate reward model. Instead, DPO directly optimizes the policy using human preference data. The DPO objective is given by:

$$\begin{aligned} rCl\mathcal{L}_{\text{DPO}}(\phi) \quad = \quad & -\mathbb{E}_{(y_1,y_2)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\phi(y_1)}{\pi_{\text{ref}}(y_1)}\right.\right.\right. \\ & \left.\left.\left. -\log\frac{\pi_\phi(y_2)}{\pi_{\text{ref}}(y_2)}\right)\right)\right] \end{aligned} \tag{7}$$

where $\sigma$ is the sigmoid function, $\mathcal{D}$ is the dataset of human preferences, and $\beta$ is a temperature parameter [10].

*13.4. Safe RLHF*

Safe RLHF introduces cost constraints to ensure that the model's outputs are both helpful and harmless. The optimization problem is formulated as:

$$\max_\phi \mathbb{E}_{y\sim\pi_\phi}[R_\theta(y)] \quad \text{subject to} \quad \mathbb{E}_{y\sim\pi_\phi}[C(y)] \le \tau, \tag{8}$$

where $C(y)$ is a cost function that measures the harmfulness of output $y$, and $\tau$ is a predefined threshold [25].

*13.5. Online Iterative RLHF*

Online iterative RLHF involves continuous updates to the policy and reward model based on new human feedback. The reward model is updated using a loss function that incorporates both historical and new preference data:

$$\begin{aligned} rCl\mathcal{L}_{\text{RM}}(\theta) \quad = \quad & -\mathbb{E}_{(y_1,y_2)\sim\mathcal{D}_{\text{new}}}[\log\sigma(R_\theta(y_1) - R_\theta(y_2))] \\ & +\lambda\,\mathrm{KL}(R_\theta||R_{\text{old}}) \end{aligned} \tag{9}$$

where $\mathcal{D}_{\text{new}}$ is the new preference data, $R_{\text{old}}$ is the previous reward model, and $\lambda$ is a regularization parameter [12].

 doi:10.20944/preprints202503.1159.v1



## 14. Pseudocode Representations of Key RLHF Algorithms

This section presents pseudocode representations of key algorithms used in Reinforcement Learning from Human Feedback (RLHF) systems, based on insights from the provided literature. The pseudocode aims to provide a high-level understanding of the algorithmic steps involved.

### *14.1. Simplified RLHF Training Loop*

```
Algorithm: Simplified RLHF Training

Input:
LLM (Language Model)
Human Feedback Data (Preference Pairs)

Output:
Fine-tuned LLM

Steps:

1. Train Reward Model (RM):
a. For each batch of Preference Pairs in Human Feedback Data:
i.   Pass both outputs to the LLM model
ii.  Compute reward score based on feedback
iii. Update RM parameters to predict higher reward for preferred output

2. Fine-tune LLM with Reinforcement Learning:
a. Initialize RL agent with LLM policy
b. For each episode:
i.   Generate text output using LLM
ii.  Use RM to predict reward for output
iii. Update LLM policy using PPO to maximize predicted reward
iv.  Implement SAFE RLHF as discussed in the SAFE RLHF paper \cite{dai_safe_2024}.
```

### *14.2. Direct Preference Optimization (DPO)*

```
Algorithm: Direct Preference Optimization (DPO)

Input:
LLM (Language Model)
Human Feedback Data (Preference Pairs)

Output:
Fine-tuned LLM

Steps:

1. Initialize LLM policy

2. For each batch of Preference Pairs in Human Feedback Data:
a.  Compute DPO loss based on preference probabilities
b.  Update LLM policy parameters to minimize DPO loss
```

### *14.3. Considerations from Web Articles and Analysis*

When building an RLHF system and integrating the chatbot discussed in [33] developers need to consider the steps and integration when doing so to minimize the impact on the user.

**Disclaimer:** These pseudocode representations are simplified for clarity and do not include all implementation details. They are intended to provide a high-level overview of the algorithms based on insights from the documents referenced.

 doi:10.20944/preprints202503.1159.v1



### 14.4. Pseudocode for Reward Model Training

**procedure** TRAINREWARDMODEL($D_{pref}$, $R_\theta$, $\theta$, $\alpha$)
  $D_{pref} \leftarrow$ Preference dataset $\{(x, y_1, y_2, \text{label})\}$
  $R_\theta \leftarrow$ Reward model with parameters $\theta$
  $\alpha \leftarrow$ Learning rate
  **while** not convergence **do**
    **for** $(x, y_1, y_2, \text{label})$ in $D_{pref}$ **do**
      $P(y_1 \succ y_2|x) \leftarrow \sigma(R_\theta(x, y_1) - R_\theta(x, y_2))$
      $\mathcal{L}(\theta) \leftarrow -\text{label} \log(P(y_1 \succ y_2|x)) - (1 - \text{label}) \log(1 - P(y_1 \succ y_2|x))$
      $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$
    **end for**
  **end while**
  **return** $R_\theta$
**end procedure**

### 14.5. Pseudocode for Proximal Policy Optimization (PPO) in RLHF

**procedure** PPO_RLHF($\pi_\phi$, $R_\theta$, $\phi$, $\epsilon$, $\alpha$)
  $\pi_\phi \leftarrow$ Policy model with parameters $\phi$
  $R_\theta \leftarrow$ Trained reward model
  $\epsilon \leftarrow$ Clipping parameter
  $\alpha \leftarrow$ Learning rate
  **while** not convergence **do**
    $D_{rollout} \leftarrow$ Collect rollout data $\{(s_t, a_t, r_t)\}$ using $\pi_\phi$
    **for** $(s_t, a_t, r_t)$ in $D_{rollout}$ **do**
      $\hat{A}_t \leftarrow$ Estimate advantage using $R_\theta$
      $r_t(\phi) \leftarrow \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{\text{old}}}(a_t|s_t)}$
      $\mathcal{L}^{\text{CLIP}}(\phi) \leftarrow \min(r_t(\phi)\hat{A}_t, \text{clip}(r_t(\phi), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)$
      $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}^{\text{CLIP}}(\phi)$
    **end for**
    $\phi_{\text{old}} \leftarrow \phi$
  **end while**
  **return** $\pi_\phi$
**end procedure**

### 14.6. Pseudocode for Direct Preference Optimization (DPO)

**procedure** DPO($\pi_\phi$, $D_{pref}$, $\phi$, $\beta$, $\alpha$)
  $\pi_\phi \leftarrow$ Policy model with parameters $\phi$
  $D_{pref} \leftarrow$ Preference dataset $\{(x, y_w, y_l)\}$
  $\beta \leftarrow$ Temperature parameter
  $\alpha \leftarrow$ Learning rate
  **while** not convergence **do**
    **for** $(x, y_w, y_l)$ in $D_{pref}$ **do**
      $\mathcal{L}^{\text{DPO}}(\phi) \leftarrow -\log\left(\frac{\exp(\pi_\phi(y_w|x)/\beta)}{\exp(\pi_\phi(y_w|x)/\beta) + \exp(\pi_\phi(y_l|x)/\beta)}\right)$
      $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}^{\text{DPO}}(\phi)$
    **end for**
  **end while**
  **return** $\pi_\phi$
**end procedure**

 doi:10.20944/preprints202503.1159.v1



## 15. Proposed Ideas and Research Proposals from the Literature

This section synthesizes proposed ideas, future research directions, and explicit research proposals identified within the provided literature on Reinforcement Learning from Human Feedback (RLHF).

### *15.1. Scaling and Efficiency*

#### 15.1.1. AI Feedback and Self-Improvement

Lee et al. [3] propose scaling RLHF using AI Feedback (RLAIF). Further research into self-improvement is needed, particularly techniques that enable LLMs to learn from their own outputs and iteratively refine their behavior.

#### 15.1.2. Efficiency in Preference Collection

The papers discuss some of the benefits of different AI integrations, but researchers could work to build more effective methods for efficient designs [35].

### *15.2. Safety and Ethical Considerations*

#### 15.2.1. Refining Safety Mechanisms

Dai et al. [25] propose safe reinforcement learning. However, more robust and reliable safety mechanisms are needed to prevent the generation of biased content.

#### 15.2.2. Addressing Reward Model Limitations

Chaudhari et al. [2] propose RLHF's potential limitations, stating that there are limitations that must be addressed when implementing new features and systems with RLHF.

### *15.3. Applications and Impact*

#### 15.3.1. Expansion to Other Domains

While Liu [33] discusses customer interfaces and chatbots, more work is needed to better integrate the system with real world solutions.

#### 15.3.2. Trinka AI

More research should be done to improve GEC [32] applications.

### *15.4. Theoretical Implications*

#### 15.4.1. Robustness and Stability

Further study of RLHF will allow for more clear research on the integration of RLHF, as stated by Wang et al. [22].

### *15.5. Enhancing Reward Model Evaluation*

**Proposal:** Develop standardized benchmarks and metrics for evaluating reward model performance beyond simple accuracy. This includes metrics for robustness, fairness, and alignment with diverse human preferences [30].

**Idea:** Implement adversarial testing to identify vulnerabilities and biases in reward models, leading to more robust and reliable reward signals.

### *15.6. Integrating AI Feedback for Scalable Alignment*

**Proposal:** Investigate hybrid RLHF-RLAIF architectures that combine human and AI feedback to achieve scalable and efficient alignment [3,11].

**Idea:** Develop automated methods for generating diverse and high-quality AI feedback, reducing the reliance on human annotators.

 doi:10.20944/preprints202503.1159.v1



### *15.7. Improving Policy Optimization Efficiency*

**Proposal:** Explore novel policy optimization algorithms that address the limitations of PPO and DPO, such as sample inefficiency and instability [10,23].

**Idea:** Investigate adaptive learning rate schedules and regularization techniques to improve the convergence and stability of RLHF training.

### *15.8. Addressing Safety and Ethical Concerns*

**Proposal:** Develop robust safety mechanisms and ethical guidelines for RLHF, including methods for detecting and mitigating harmful biases and unintended behaviors [25].

**Idea:** Incorporate human value alignment into the reward model design to ensure that LLMs adhere to ethical principles and societal norms.

### *15.9. Exploring Theoretical Foundations*

**Proposal:** Conduct further theoretical analysis of RLHF to understand its fundamental limitations and challenges compared to standard RL [4,22].

**Idea:** Develop formal models and frameworks for analyzing the convergence and sample complexity of RLHF algorithms.

### *15.10. Advancing Online Iterative RLHF*

**Proposal:** Investigate techniques for improving the efficiency and robustness of online iterative RLHF, including methods for handling noisy and sparse human feedback [12].

**Idea:** Develop adaptive strategies for dynamically adjusting the reward model and policy based on the quality and quantity of incoming feedback.

### *15.11. Comprehensive Surveys and Tutorials*

**Proposal:** Continue to produce comprehensive surveys and tutorials that synthesize the latest advancements in RLHF and related areas, facilitating knowledge dissemination and research collaboration [37,39,40].

**Idea:** Develop interactive educational tools and platforms that enable researchers and practitioners to explore and experiment with RLHF algorithms.

### *15.12. Other Proposed Ideas*

RLHF has expanded in recent times as described by Lambert [31] and has been summarized from other articles and studies online [1]. More research is needed to summarize and share this information with the world. This section consolidates proposed ideas and research proposals derived from the key references, highlighting potential avenues for future exploration and development in RLHF.

### *15.13. Cloud Platform Integration for RLHF*

Cloud platforms play a crucial role in enabling scalable and efficient RLHF implementations. This section explores the offerings of major cloud providers, including Google Cloud and Amazon Web Services (AWS), and their related products for RLHF, based on the provided keys.

#### 15.13.1. Google Cloud

Google Cloud provides resources to tune models like PaLM 2 and Llama 2 with RLHF [27].

#### 15.13.2. Amazon Web Services (AWS)

AWS offers a range of services that support RLHF workflows. Specifically, AWS provides information on improving LLMs with RLHF on Amazon SageMaker [26].

 doi:10.20944/preprints202503.1159.v1



15.13.3. Infrastructure and Scalability

GitHub documentation and web articles highlight [35] discuss essential tools for the process. Tools like the ones discussed on GitHub and through the papers are essential components that cloud platforms such as Google Cloud or Amazon Web Services provides. These services enable model training and deployment on large datasets. Cloud platforms may allow for improved customer services [33].

## 16. Future Directions and Conclusion

Future research in RLHF should focus on addressing the challenges and limitations discussed in this paper. One promising direction is the development of more efficient feedback collection methods, such as the use of groupwise comparisons and interactive feedback mechanisms [24]. Another important area of research is the improvement of reward model generalization, which could be achieved through the use of contrastive learning and meta-learning [29].

In addition to these technical challenges, future research should also focus on the ethical implications of RLHF. This includes the development of transparent and accountable feedback collection processes, as well as methods for ensuring the privacy and consent of human annotators [4].

Future research directions include exploring more efficient and scalable RLHF methods, addressing the challenges of data collection and annotation, and improving the robustness and safety of LLMs. Additionally exploring RLAIF is important [3,11]. Surveys of LLM alignment techniques, including RLHF, RLAIF, PPO, and DPO, are also crucial [37]. Tutorials on LLM reasoning and hallucination are also important [39,40].

RLHF is a powerful technique for aligning LLMs with human preferences. Continued research and development in this area will contribute to the creation of more reliable, safe, and beneficial AI systems.

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful technique for aligning large language models with human preferences. This paper has provided a comprehensive review of RLHF, covering its theoretical foundations, practical implementations, and recent advancements. We have also discussed the challenges and limitations of current approaches, including the reliance on human feedback, the generalization of reward models, and the ethical implications of RLHF.

Future research should focus on addressing these challenges, including the development of more efficient feedback collection methods, the improvement of reward model generalization, and the ethical use of RLHF. By addressing these issues, RLHF has the potential to play a key role in the development of safe and effective AI systems.

## References

1. "What Is Reinforcement Learning from Human Feedback (RLHF)?" [Online]. Available: https://swimm.io/learn/large-language-models/what-is-reinforcement-learning-from-human-feedback-rlhf
2. S. Chaudhari, P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, A. Deshpande, and B. C. da Silva, "RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs," 2024, publisher: arXiv Version Number: 2. [Online]. Available: https://arxiv.org/abs/2404.08555
3. H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash, "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback."
4. S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell, "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," Sep. 2023, arXiv:2307.15217 [cs]. [Online]. Available: http://arxiv.org/abs/2307.15217
5. "Aligning language models to follow instructions," Feb. 2024. [Online]. Available: https://openai.com/index/instruction-following/

 doi:10.20944/preprints202503.1159.v1




6. "Learning from human preferences," Feb. 2024. [Online]. Available: https://openai.com/index/learning-from-human-preferences/
7. "Reinforcement Learning from Human Feedback (RLHF) Explained." [Online]. Available: https://mediacenter.ibm.com/media/Reinforcement+Learning+from+Human+Feedback+%28RLHF%29+Explained/1_uv1w3sj3
8. "What Is Reinforcement Learning From Human Feedback (RLHF)? | IBM," Nov. 2023. [Online]. Available: https://www.ibm.com/think/topics/rlhf
9. "What is RLHF? - Reinforcement Learning from Human Feedback Explained - AWS." [Online]. Available: https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/
10. R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, Dec. 2023. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2023
11. D. Mahan, D. V. Phung, R. Rafailov, C. Blagden, N. Lile, L. Castricato, J.-P. Fränken, C. Finn, and A. Albalak, "Generative Reward Models - A Unified Approach to RLHF and RLAIF."
12. H. Dong, W. Xiong, B. Pang, H. Wang, H. Zhao, Y. Zhou, N. Jiang, D. Sahoo, C. Xiong, and T. Zhang, "RLHF Workflow: From Reward Modeling to Online RLHF," Nov. 2024, arXiv:2405.07863 [cs]. [Online]. Available: http://arxiv.org/abs/2405.07863
13. Satyadhar Joshi, "Advancing Financial Risk Modeling: Vasicek Framework Enhanced by Agentic Generative AI by Satyadhar Joshi," *Advancing Financial Risk Modeling: Vasicek Framework Enhanced by Agentic Generative AI by Satyadhar Joshi*, vol. Volume 7, no. Issue 1, January 2025, Jan. 2025.
14. S. Joshi, "Advancing innovation in financial stability: A comprehensive review of ai agent frameworks, challenges and applications," *World Journal of Advanced Engineering Technology and Sciences*, vol. 14, no. 2, pp. 117–126, 2025.
15. ——, *Agentic Gen AI For Financial Risk Management*. Draft2Digital, 2025.
16. Satyadhar Joshi, "Enhancing structured finance risk models (Leland-Toft and Box-Cox) using GenAI (VAEs GANs)," *International Journal of Science and Research Archive*, vol. 14, no. 1, pp. 1618–1630, 2025.
17. S. Joshi, "Implementing Gen AI for Increasing Robustness of US Financial and Regulatory System," *International Journal of Innovative Research in Engineering and Management*, vol. 11, no. 6, pp. 175–179, Jan. 2025.
18. Satyadhar Joshi, "Review of Gen AI Models for Financial Risk Management," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 709–723, Jan. 2025.
19. Satyadhar Joshi, "The Synergy of Generative AI and Big Data for Financial Risk: Review of Recent Developments," *IJFMR - International Journal For Multidisciplinary Research*, vol. 7, no. 1.
20. Satyadhar Joshi, "Leveraging prompt engineering to enhance financial market integrity and risk management," *World Journal of Advanced Research and Reviews*, vol. 25, no. 1, pp. 1775–1785, Jan. 2025.
21. "(PDF) 7 Reinforcement Learning from Human Feedback (RLHF)." [Online]. Available: https://www.researchgate.net/publication/383931211_7_Reinforcement_Learning_from_Human_Feedback_RLHF
22. Y. Wang, Q. Liu, and C. Jin, "Is RLHF More Difficult than Standard RL? A Theoretical Perspective."
23. R. Zheng, S. Dou, S. Gao, Y. Hua, W. Shen, B. Wang, Y. Liu, S. Jin, Q. Liu, Y. Zhou, L. Xiong, L. Chen, Z. Xi, N. Xu, W. Lai, M. Zhu, C. Chang, Z. Yin, R. Weng, W. Cheng, H. Huang, T. Sun, H. Yan, T. Gui, Q. Zhang, X. Qiu, and X. Huang, "Secrets of RLHF in Large Language Models Part I: PPO."
24. J. Kompatscher, "Interactive Groupwise Comparison for Faster Reinforcement Learning from Human Feedback," Dec. 2024. [Online]. Available: https://aaltodoc.aalto.fi/handle/123456789/133612
25. J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "SAFE RLHF: SAFE REINFORCEMENT LEARNING FROM HUMAN FEEDBACK," 2024.
26. "Improving your LLMs with RLHF on Amazon SageMaker | AWS Machine Learning Blog," Sep. 2023, section: Amazon SageMaker. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/improving-your-llms-with-rlhf-on-amazon-sagemaker/
27. "RLHF on Google Cloud." [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/rlhf-on-google-cloud
28. "RLHF - Hugging Face Deep RL Course." [Online]. Available: https://huggingface.co/learn/deep-rl-course/en/unitbonus3/rlhf
29. B. Wang, R. Zheng, L. Chen, Y. Liu, S. Dou, C. Huang, W. Shen, S. Jin, E. Zhou, C. Shi, S. Gao, N. Xu, Y. Zhou, X. Fan, Z. Xi, J. Zhao, X. Wang, T. Ji, H. Yan, L. Shen, Z. Chen, T. Gui, Q. Zhang, X. Qiu, X. Huang,

 doi:10.20944/preprints202503.1159.v1



Z. Wu, and Y.-G. Jiang, "Secrets of RLHF in Large Language Models Part II: Reward Modeling," Jan. 2024, arXiv:2401.06080 [cs]. [Online]. Available: http://arxiv.org/abs/2401.06080
30. E. Frick, T. Li, C. Chen, W.-L. Chiang, A. N. Angelopoulos, J. Jiao, B. Zhu, J. E. Gonzalez, and I. Stoica, "How to Evaluate Reward Models for RLHF," Oct. 2024, arXiv:2410.14872 [cs]. [Online]. Available: http://arxiv.org/abs/2410.14872
31. N. Lambert, "(WIP) A Little Bit of Reinforcement Learning from Human Feedback."
32. "RLHF for Grammar Error Correction Trinka," Dec. 2024, section: Technical Blog. [Online]. Available: https://www.trinka.ai/blog/rlhf-for-grammar-error-correction/
33. J. Liu, "ChatGPT: perspectives from human–computer interaction and psychology," *Frontiers in Artificial Intelligence*, vol. 7, Jun. 2024, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1418869/full
34. A. Briouya, H. Briouya, and A. Choukri, "Overview of the progression of state-of-the-art language models," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 4, pp. 897–909, Aug. 2024, number: 4. [Online]. Available: https://telkomnika.uad.ac.id/index.php/TELKOMNIKA/article/view/25936
35. "RLHF/main.pdf at main · Peymankor/RLHF." [Online]. Available: https://github.com/Peymankor/RLHF/blob/main/main.pdf
36. Y. Xu, T. Chakraborty, E. Kıcıman, B. Aryal, E. Rodrigues, S. Sharma, R. Estevao, M. A. d. L. Balaguer, J. Wolk, R. Padilha, L. Nunes, S. Balakrishnan, S. Lu, and R. Chandra, "RLTHF: Targeted Human Feedback for LLM Alignment," Feb. 2025, arXiv:2502.13417 [cs]. [Online]. Available: http://arxiv.org/abs/2502.13417
37. Z. Wang, B. Bi, S. K. Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, Zixu, Zhu, X.-B. Mao, S. Asur, Na, and Cheng, "A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More," Jul. 2024, arXiv:2407.16216 [cs]. [Online]. Available: http://arxiv.org/abs/2407.16216
38. G. Li, X. Zhou, and X. Zhao, "LLM for Data Management," *Proc. VLDB Endow.*, vol. 17, no. 12, pp. 4213–4216, Aug. 2024. [Online]. Available: https://doi.org/10.14778/3685800.3685838
39. J. Wang, "A Tutorial on LLM Reasoning: Relevant Methods behind ChatGPT o1," Feb. 2025, arXiv:2502.10867 [cs]. [Online]. Available: http://arxiv.org/abs/2502.10867
40. V. Rawte, A. Chadha, A. Sheth, and A. Das, "Tutorial Proposal: Hallucination in Large Language Models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, R. Klinger, N. Okazaki, N. Calzolari, and M.-Y. Kan, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 68–72. [Online]. Available: https://aclanthology.org/2024.lrec-tutorials.11/