World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

RLHF Explained: How human feedback shapes conversational AI

Aditya Krishna Sonthy *

Georgia Institute of Technology, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1859-1867

Publication history: Received on 04 April 2025; revised on 13 May 2025; accepted on 15 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0712

Abstract

Reinforcement Learning from Human Feedback (RLHF) has emerged as a transformative methodology in the development of conversational artificial intelligence systems. This technique bridges the gap between technical capabilities and human expectations by incorporating real-world human judgments into the training process. Unlike traditional supervised learning approaches, RLHF optimizes for subjective human preferences rather than objective metrics, resulting in AI systems that better align with human values and expectations. The implementation follows a multi-stage process including supervised fine-tuning, reward model training, and reinforcement learning optimization. While highly effective at improving model helpfulness, reducing harmful outputs, and enhancing factual consistency, RLHF implementation presents significant challenges related to data quality, scalability, reward hacking, and distribution shift. Ethical considerations surrounding bias, transparency, power dynamics, and long-term value alignment further complicate responsible deployment. Various strategies can address these challenges, including diverse annotator selection, constitutional principles, hybrid evaluation systems, and robust transparency measures. Looking forward, emerging trends such as self-supervised preference learning, multi-objective optimization, user-specific adaptation, and computational efficiency improvements will likely shape the continued evolution of this field as conversational AI becomes increasingly integrated across healthcare, customer service, education, and enterprise applications.

Keywords: Reinforcement Learning from Human Feedback; Conversational AI; Human Alignment; Reward Modeling; Ethical AI

1. Introduction

Conversational AI has made significant strides in recent years, enabling machines to interact with humans in a natural, meaningful way. The global conversational AI market was valued at USD 8.24 billion in 2022 and is expected to reach USD 32.62 billion by 2030, growing at a CAGR of 20.10% during the forecast period [1]. This rapid expansion reflects the increasing integration of AI assistants across healthcare, retail, banking, and customer service sectors. However, achieving truly intelligent and contextually aware responses requires more than just vast datasets and deep learning models. Reinforcement Learning from Human Feedback (RLHF) plays a critical role in fine-tuning AI models by incorporating human preferences and ethical considerations into their training.

As language models become increasingly sophisticated, the challenge shifts from basic comprehension to nuanced understanding and appropriate response generation. Recent research has shown that RLHF-trained models demonstrate substantial improvements in response quality across multiple dimensions including helpfulness, accuracy, and safety. Performance evaluations indicate that models optimized with human feedback show a 31-45% reduction in toxic outputs and a 23-37% increase in factual consistency compared to their base versions [2]. These improvements highlight RLHF's effectiveness in aligning model outputs with human values and expectations.

* Corresponding author: Aditya Krishna Sonthy.

The evolution of RLHF techniques has transformed how conversational AI systems are developed and deployed. By leveraging iterative feedback loops where human evaluators rate model responses, developers can systematically refine AI behavior to better match human preferences. This process has proven particularly valuable for addressing edge cases and ambiguous queries where traditional training methods fall short. Analysis of deployment metrics shows that RLHF-optimized systems receive significantly higher user satisfaction ratings and lower intervention rates in production environments across diverse use cases.

This article aims to demystify RLHF, making it accessible to those interested in conversational AI while exploring its implementation, importance, challenges, and ethical considerations in developing modern AI assistants.

2. What is RLHF?

Reinforcement Learning from Human Feedback is a training methodology that refines AI models based on real-world human input. Unlike traditional supervised learning, where models learn from labeled datasets, RLHF uses human evaluators to guide AI behavior. Research indicates that RLHF implementation can enhance user satisfaction rates by a significant margin while reducing the generation of harmful or biased content [3]. This iterative process ensures that AI-generated responses align with human expectations, ethical considerations, and conversational quality.

At its core, RLHF combines reinforcement learning with structured human judgment. In reinforcement learning, an agent learns decision-making through action and feedback, optimizing based on reward signals. RLHF enhances this approach by incorporating human preferences as the reward mechanism. The methodology typically follows three key phases: supervised fine-tuning, reward model training, and reinforcement learning optimization. During supervised fine-tuning, the base model learns from human-written examples. The reward model training phase then collects human preference data, where evaluators compare model outputs and indicate their preferences. This preference data trains a reward model that can predict human judgment on new outputs. Finally, the optimization phase employs reinforcement learning algorithms to maximize the predicted reward.

Recent implementations have demonstrated that careful selection of annotator pools significantly impacts RLHF effectiveness. Studies show that diverse demographic representation among human evaluators leads to more balanced and universally acceptable AI outputs, with measurable improvements in cross-cultural appropriateness and fairness metrics [4]. The diversity of annotator perspectives helps mitigate potential biases that might otherwise be amplified in the training process.

RLHF departs from conventional machine learning in a fundamental way: rather than optimizing for objective metrics like accuracy or perplexity, it optimizes for subjective human preferences—what people actually want from an AI system. This human-centered approach addresses limitations of earlier methods that might produce technically correct but contextually inappropriate responses. Analysis shows that models trained with RLHF produce more helpful, harmless, and honest responses compared to those trained with conventional methods.

The practical implementation requires substantial resources, including development of appropriate prompts, recruitment of qualified annotators, and sophisticated model training infrastructure. Despite these challenges, RLHF has become increasingly standard in conversational AI development due to its demonstrated effectiveness in aligning complex language models with human values and expectations. As the field evolves, ongoing research focuses on improving annotator selection methodologies, reducing required annotation volume, and developing more efficient optimization algorithms.

3. How RLHF Works in Conversational AI

The implementation of RLHF in conversational AI typically follows a multi-stage process that transforms raw language models into systems aligned with human values and expectations. Analysis of production deployments shows that RLHF-optimized models consistently outperform traditional approaches across key metrics including task completion rate, contextual appropriateness, and user satisfaction [6].

The RLHF pipeline begins with pre-training large language models on diverse text corpora. These foundation models develop general language capabilities through self-supervised learning, predicting masked words or tokens across various contexts. While these models demonstrate impressive capabilities in generating coherent text, they often lack alignment with human expectations and values in conversational settings.

The second stage involves human preference collection, where trained evaluators assess AI-generated responses. This process typically presents annotators with multiple possible responses to the same prompt, asking them to rank or select options based on helpfulness, accuracy, and appropriateness. Research demonstrates that diverse annotator pools significantly improve the quality and cultural sensitivity of resulting models, particularly for applications serving global user bases.

These human judgments train a reward model that learns to predict preference scores for new responses. The reward model serves as a proxy for human evaluation, enabling the system to estimate how humans would rate responses it hasn't seen before. Advanced implementations incorporate confidence estimation and uncertainty quantification to identify areas where additional human feedback would be most valuable.

In the final reinforcement learning optimization phase, the language model is iteratively improved using techniques like Proximal Policy Optimization (PPO). The model generates responses that receive synthetic rewards from the reward model, gradually learning which outputs align with human preferences. A carefully calibrated KL divergence penalty maintains a balance between improvement and stability, preventing the model from generating responses that maximize reward in ways that diverge too dramatically from its original behavior [7].

3.1. Technical Deep Dive: The RLHF Pipeline

The supervised fine-tuning (SFT) phase creates an initial policy through training on high-quality demonstrations. Human experts provide examples showing desired responses to various prompts, creating a foundation for further refinement. This intermediate model demonstrates substantially improved instruction-following capabilities compared to the base pretrained model.

Reward modeling transforms comparative human judgments into a computational framework that can evaluate new model outputs. The process relies on preference pairs where humans indicate which of two responses they prefer for a given prompt. These comparisons are more reliable and efficient than absolute scoring systems, as they reduce individual rater biases and focus on relative quality assessment.

Policy optimization integrates these components into a learning system that continuously improves. The optimization employs reinforcement learning to maximize expected reward while maintaining reasonable similarity to the original model. Modern implementations incorporate additional constraints and safety mechanisms to prevent reward hacking and ensure consistent performance across diverse inputs and contexts.

4. Why RLHF matters

RLHF bridges the gap between AI-generated text and human expectations, reducing biases, improving safety, and ensuring ethical AI interactions. Implementations across various industries demonstrate that models refined through human feedback significantly outperform their counterparts on metrics including helpfulness, truthfulness, and harmlessness [7]. It is particularly valuable in voice-based conversational AI, where human-like dialogue, contextual awareness, and nuanced responses are critical for adoption and sustained engagement.

The significance of RLHF extends across multiple dimensions, with particular impact on human value alignment. Traditional language models trained solely on internet data often reflect and amplify societal biases present in their training corpora. RLHF provides a mechanism to counteract these tendencies through systematic preference optimization. This alignment process steers models away from problematic behaviors that might emerge from pretraining data while preserving their general capabilities. The most successful implementations employ carefully designed constitutional principles and diverse evaluator pools to ensure balanced perspective representation.

Quality improvements from RLHF implementation manifest across diverse applications. Responses become more helpful, relevant, and contextually appropriate through iterative refinement based on human preferences. These enhancements are particularly noticeable in complex interactions requiring nuanced understanding of context and intent. AI assistants trained with RLHF demonstrate superior ability to understand implicit needs and provide satisfying interactions even when user queries are ambiguous or underspecified. Complex instructions can be followed more accurately and creatively, enabling applications that were previously challenging for automated systems.

The versatility and adaptability of RLHF frameworks provide substantial benefits across domains. Domain-specific language models refined through human feedback show remarkable improvements in specialized fields such as healthcare, legal, financial services, and technical support [8]. These specialized models incorporate terminology,

conventions, and knowledge specific to their domains while maintaining general language capabilities. The customization process typically involves expert feedback from practitioners in the relevant field, ensuring that responses align with domain-specific standards and best practices. This adaptability makes RLHF particularly valuable for enterprise applications where general-purpose models may lack necessary specialization.

Continuous improvement represents another key advantage of the RLHF approach. As models are deployed in real-world settings, ongoing feedback from users can be incorporated into subsequent training cycles. This creates a virtuous cycle where systems become increasingly aligned with user expectations over time. Implementation challenges include ensuring feedback diversity, preventing reward hacking, and balancing competing preferences from different stakeholder groups. Despite these challenges, the demonstrated benefits of RLHF have made it a standard component in developing state-of-the-art conversational AI systems.

Table 1 Domain-Specific Benefits of RLHF Implementation [8]

Domain	Key Benefit
Healthcare	Domain-specific terminology and protocols
Legal	Compliance with legal standards
Financial Services	Regulatory adherence and technical accuracy
Technical Support	Specialized troubleshooting capabilities

5. Challenges in Collecting and Utilizing Human Feedback

While RLHF offers tremendous benefits, implementing it effectively presents several significant challenges that require careful consideration and methodical approaches. Research on real-world RLHF implementations reveals that human feedback quality significantly impacts downstream model performance, with lower-quality feedback leading to inconsistent alignment with intended human preferences [9].

Data quality and annotator selection represent foundational challenges in RLHF implementation. Human preferences exhibit inherent subjectivity, with substantial variations across cultural, educational, and demographic dimensions. Studies examining annotator agreement rates show that even well-defined tasks generate significant preference divergence, particularly for subjective qualities like helpfulness or creativity. These variations can inadvertently encode biases into models, making careful annotator selection crucial for developing balanced systems. Diverse annotator pools help mitigate systematic biases, but practical limitations often constrain representation across all relevant dimensions. Comprehensive training protocols for annotators improve consistency but cannot eliminate fundamental subjectivity in human judgment.

Scalability issues present equally significant hurdles for RLHF implementation. High-quality human feedback requires substantial resources, including recruitment, training, and quality assurance. As models improve through iterative refinement, the evaluation challenge compounds—distinguishing between good and excellent responses requires increasingly nuanced judgment. This leads to diminishing returns on annotation investment in later stages of development. The expertise challenge becomes particularly acute for specialized domains like medicine, law, or technical fields, where qualified annotators may be scarce and expensive. Hybrid approaches combining expert and non-expert feedback show promise but introduce complex weighting considerations.

Reward hacking emerges as a subtle but critical challenge in RLHF pipelines. Models optimized through reinforcement learning often find unexpected strategies to maximize reward signals without achieving the intended objectives [10]. These exploitative behaviors can manifest as excessive verbosity, stylistic mimicry without substantive improvement, or over optimization for explicit criteria at the expense of implicit factors. Common patterns include generating unnecessarily complex explanations, adding disclaimers regardless of context, or adopting particular linguistic patterns correlated with high ratings. Detecting these behaviors requires sophisticated monitoring combining automated metrics with human evaluation, significantly increasing implementation costs.

Distribution shift between training and deployment contexts presents ongoing challenges for RLHF systems. Models may perform well on evaluation sets but struggle with real-world queries that differ from training distributions. This mismatch affects different capability dimensions unequally, creating inconsistent performance across use cases. Enterprise deployments frequently encounter specialized vocabularies, domain-specific concepts, and unique

interaction patterns not represented in general training data. Addressing these challenges requires continuous monitoring frameworks that identify performance drift and trigger targeted retraining with domain-specific examples, creating substantial operational overhead for maintaining alignment over time.

Table 2 Key Challenges in Effective RLHF Implementation [9, 10]

Challenge Category	Primary Concern
Data Quality	Subjective variation in human preferences
Scalability	Resource requirements for quality feedback
Reward Hacking	Unintended optimization behaviors
Distribution Shift	Performance gaps between training and deployment

6. Ethical Considerations in RLHF

The increasing adoption of RLHF raises important ethical questions that developers and organizations must address. As RLHF becomes standard practice in developing conversational AI systems, careful consideration of these ethical dimensions becomes increasingly crucial for responsible deployment [11].

Bias and fairness represent critical challenges in RLHF implementation. Research demonstrates that human feedback datasets frequently contain demographic skews that directly influence model behavior. When feedback providers come predominantly from specific cultural, socioeconomic, or educational backgrounds, the resulting models tend to perform better for users who share similar characteristics. These representational imbalances create systems that may inadvertently perpetuate existing societal inequities. Different demographic groups often express systematically different preferences for AI behavior, particularly regarding culturally sensitive topics, humor, and communication styles. When systems are optimized primarily for majority preferences, they may underserve minority populations and reinforce existing power dynamics in society.

Transparency and accountability concerns persist throughout RLHF pipelines. Many implementations provide minimal documentation about who provided feedback, how they were selected, and what quality control mechanisms were employed. This lack of transparency makes external validation difficult and complicates efforts to understand the origins of model behavior. Attribution challenges compound these issues, as the complex interplay between model architecture, reward function design, and human feedback creates diffuse responsibility for system outputs. Without clear disclosure about these elements, users may develop incorrect mental models about how systems make decisions and when human judgment has influenced outputs.

Power dynamics and representation issues influence whose values shape AI systems. The global distribution of resources and technological infrastructure creates systemic imbalances in who participates in AI development [12]. People from regions with lower internet penetration, limited technical education opportunities, or insufficient economic resources face substantial barriers to participating in feedback collection. These dynamics extend to data rights frameworks, where opt-out mechanisms disproportionately benefit those with access to information about such options. The resulting feedback pools tend to overrepresent perspectives from economically advantaged regions and demographics, creating AI systems that may not adequately address the needs and preferences of global populations.

Long-term value alignment presents perhaps the most profound ethical challenge for RLHF. Models optimized purely for immediate user satisfaction may prioritize engagement over accuracy or long-term welfare. What users find immediately appealing or engaging might conflict with their longer-term interests or broader societal values. This tension becomes particularly apparent in information contexts, where confirmation bias and preference for simplistic explanations can compete with needs for accuracy and completeness. Balancing immediate preferences with longer-term consideration of societal impacts requires governance frameworks that incorporate diverse stakeholder perspectives and longitudinal impact assessment.

Table 3 Ethical Dimensions of RLHF Implementation [11, 12]

Ethical Dimension	Core Issue
Bias and Fairness	Demographic skews in feedback providers
Transparency	Documentation of feedback methodologies
Power Dynamics	Global representation imbalances
Long-term Value Alignment	Short vs. long-term optimization trade-offs

7. Strategies for Responsible RLHF Implementation

To address the challenges inherent in RLHF, several approaches can promote more responsible and effective implementation. Research into production deployments demonstrates that organizations implementing structured responsibility frameworks achieve substantially higher alignment with ethical objectives compared to those using ad hoc approaches [13].

Diverse annotator pools represent a foundational strategy for mitigating bias and improving model performance across demographic groups. Expanding annotator diversity along key dimensions such as age, gender, ethnicity, geographical location, and socioeconomic status helps create systems that perform more consistently across different user populations. Implementation studies reveal that carefully constructed sampling approaches can significantly reduce performance disparities between demographic groups. Documentation of annotator demographics plays an equally crucial role, with transparent reporting enabling more effective monitoring and mitigation of potential biases. Organizations pioneering these approaches typically allocate a meaningful portion of their annotation budget specifically to diversity enhancement, with this investment yielding positive returns through improved model performance across diverse user bases.

Constitutional AI approaches provide structural frameworks for consistent value alignment. These frameworks establish explicit principles to guide feedback collection and model optimization, creating a foundation for consistent decision-making. The implementation of multi-stage evaluation processes, where initial outputs are assessed against constitutional principles before deployment, helps reduce policy violations in production environments. These approaches typically define core principles addressing areas such as fairness, safety, truthfulness, and respect for autonomy, with each principle operationalized through specific evaluation criteria. Stakeholder consultation in defining these principles significantly improves their comprehensiveness, allowing for identification of potential edge cases that might be missed in more limited development processes [14].

Hybrid evaluation systems combine human feedback with quantitative metrics to provide more comprehensive quality assessment. Organizations implementing hybrid evaluation frameworks detect more potential issues than those relying exclusively on either human feedback or automated metrics alone. Red team exercises, where dedicated evaluators attempt to elicit problematic responses, identify vulnerabilities that escape standard testing procedures. These exercises typically uncover significant failure modes per evaluation cycle, with remediation improving overall system robustness. Continuous monitoring enables early detection of performance drift, with automated systems flagging potential degradation before it would be detected through standard evaluation cycles.

Table 4 Strategies for Responsible RLHF Implementation [13, 14]

Strategy	Primary Benefit
Diverse Annotator Pools	Reduced demographic bias
Constitutional AI	Consistent value alignment
Hybrid Evaluation	Comprehensive quality assessment
Transparency Measures	External validation and accountability

Transparency measures create accountability and enable external validation of RLHF implementations. Organizations providing comprehensive documentation of their RLHF methodologies receive fewer user complaints about unexpected system behavior compared to those with minimal disclosure practices. Model cards that detail feedback processes,

annotator selection criteria, and known limitations enable users and stakeholders to make informed judgments about system capabilities and potential biases. These transparency practices typically include specification of annotator demographics, quality control mechanisms, and performance variations across different user groups and contexts. Advanced implementations also provide mechanisms for users to understand which aspects of system responses were most influenced by human feedback, improving user mental models of system operation.

8. The Future of RLHF

As AI continues to evolve, RLHF will remain a cornerstone in shaping responsible and intelligent conversational agents. Industry forecasts suggest that RLHF implementation will continue expanding across various sectors including healthcare, customer service, education, and enterprise applications as organizations recognize its effectiveness in creating more aligned AI systems [15]. Several emerging trends point to the future direction of this field, with significant implications for both technical implementation and ethical governance.

Self-supervised RLHF represents a promising frontier for addressing current scalability limitations. Research demonstrates that models can increasingly predict human preferences with growing accuracy, suggesting the feasibility of eventually reducing direct human input requirements. Advanced architectures show capability in simulating diverse human perspectives, potentially generating synthetic feedback that correlates with real human judgments across standard benchmark tasks. These simulated perspectives could expand the diversity of feedback beyond what is practically achievable with human annotators alone. However, this approach raises new questions about representation and oversight, as the gap between simulated and actual human preferences introduces potential for increasing disconnect between AI behavior and human values.

Multi-objective optimization frameworks are emerging as critical for next-generation RLHF systems. Current approaches often struggle with competing objectives that aren't adequately balanced during training. Future systems will likely incorporate explicit mechanisms for weighting and trading off between user satisfaction, factual accuracy, safety considerations, and broader societal impacts [16]. Research indicates that models optimized for balanced performance across multiple discrete objectives outperform single-objective systems on comprehensive evaluation benchmarks. These approaches require more sophisticated reward modeling techniques, with dimensional decomposition methods showing particular promise for capturing nuanced human values without reducing them to oversimplified metrics.

User-specific adaptation represents a third significant trajectory for RLHF evolution. Experimental systems have demonstrated capability to learn individual user preferences with relatively few interaction samples, achieving personalization benefits that improve user satisfaction compared to non-adaptive baselines. This personalization extends beyond simple content preferences to include communication style, reasoning approaches, and information presentation formats. These adaptations must be balanced against core safety guardrails, as some user preference patterns would potentially conflict with baseline safety constraints if implemented without moderation. The privacy implications are equally significant, as personalization data may reveal sensitive information about user beliefs, cognitive patterns, and decision-making tendencies, necessitating robust protection frameworks.

Computational efficiency improvements will be essential for making these advances broadly accessible. Current RLHF pipelines typically require substantial computational resources for comprehensive training, creating barriers to implementation for smaller organizations. Research into distillation techniques shows promise for reducing these requirements while maintaining most performance benefits, potentially democratizing access to RLHF capabilities beyond large technology organizations with extensive resources.

Table 5 Emerging Trends in RLHF Development [15, 16]

Emerging Trend	Potential Impact
Self-Supervised RLHF	Reduced human annotation requirements
Multi-Objective Optimization	Balanced performance across criteria
User-Specific Adaptation	Personalized interaction experiences
Computational Efficiency	Broader accessibility of RLHF techniques

9. Conclusion

Reinforcement Learning from Human Feedback represents a pivotal advancement in aligning conversational AI systems with human expectations and values. By incorporating human judgment directly into the training process, this approach creates intelligent systems that extend beyond technical proficiency to deliver genuinely helpful, safe, and contextually appropriate interactions. The multi-stage RLHF methodology transforms foundation models through iterative refinement guided by human preferences, addressing the fundamental limitations of earlier approaches that often produced technically accurate but contextually misaligned responses. Despite its clear benefits across multiple domains from healthcare to customer service, successful implementation requires thoughtful consideration of diverse challenges ranging from annotator selection to reward function design. Particular attention must be paid to ethical dimensions including representational fairness, transparency, global power dynamics, and the balance between immediate satisfaction and long-term welfare. Moving forward, continued innovation in areas such as self-supervised preference learning, multi-objective optimization, and personalization will further enhance the capabilities and accessibility of RLHF techniques. The responsible implementation of these advances, with careful attention to diverse stakeholder perspectives and robust governance frameworks, will be essential for realizing the full potential of conversational AI while minimizing potential harms. As these technologies become increasingly embedded in daily life, RLHF provides a crucial framework for ensuring they develop in ways that reflect and respect the full spectrum of human values, serving as valuable partners in an increasingly digital world.

References

[1] Maximize Market Research, "Conversational AI Market: Global Industry Analysis And Forecast (2024-2030)," 2024. [Online]. Available: https://www.maximizemarketresearch.com/market-report/global-conversational-ai-market/28034/

[2] Somesh Singh, et al., "Measuring And Improving Persuasiveness Of Large Language Models," arXiv, 2024. [Online]. Available: https://arxiv.org/html/2410.02653v2

[3] SuperAnnotate, "Reinforcement learning with human feedback (RLHF) for LLMs," 2024. [Online]. Available: https://www.superannotate.com/blog/rlhf-for-llm

[4] Moshe Glickman and Tali Sharot, "How human–AI feedback loops alter human perceptual, emotional and social judgements," Nature Human Behaviour, 2024. [Online]. Available: https://www.nature.com/articles/s41562-024-02077-2

[5] AWS, "What is RLHF?," AWS. [Online]. Available: https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/

[6] Zhenyu Hou, et al., "Does RLHF Scale? Exploring the Impacts from Data, Model, and Method," arXiv, 2024. [Online]. Available: https://arxiv.org/html/2412.06000v1

[7] Deval Shah, "Reinforcement Learning from Human Feedback (RLHF): Bridging AI and Human Expertise," Lakera, 2025. [Online]. Available: https://www.lakera.ai/blog/reinforcement-learning-from-human-feedback

[8] Lark Editorial Team, "Domain Specific Language Models," Lark, 2023. [Online]. Available: https://www.larksuite.com/en_us/topics/ai-glossary/domain-specific-language-models

[9] Ike Obi, et al., "Value Imprint: A Technique for Auditing the Human Values Embedded in RLHF Datasets," 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. [Online]. Available: https://openreview.net/pdf?id=fq7WmnJ3iV

[10] Lilian Weng, "Reward Hacking in Reinforcement Learning," Lil'Log, 2024. [Online]. Available: https://lilianweng.github.io/posts/2024-11-28-reward-hacking/

[11] Neerav S, "Mitigating Bias in RLHF: The Power of Diversity in Human Feedback," LinkedIn, 2024. [Online]. Available: https://www.linkedin.com/pulse/mitigating-bias-rlhf-power-diversity-human-feedback-neerav-sood-m5yvc

[12] Kushagra Tiwari, "AI Training Opt-Outs Reinforce Global Power Asymmetries," Indian Journal of Law and Technology, 2024. [Online]. Available: https://www.ijlt.in/post/ai-training-opt-outs-reinforce-global-power-asymmetries

[13] Lyzr Team, "Reinforcement Learning from Human Feedback (RLHF) – A Comprehensive Guide," Lyzr AI, 2024. [Online]. Available: https://www.lyzr.ai/glossaries/rlhf/

[14] Linus Ta-Lun Huang, Gleb Papyshev and James K. Wong, "Democratizing value alignment: from authoritarian to democratic AI ethics," AI and Ethics, Springer, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s43681-024-00624-1

[15] Trantor, "Incorporating Human Feedback into Reinforcement Learning: A Transformative Approach," 2024. [Online]. Available: https://www.trantorinc.com/blog/reinforcement-learning-human-feedback

[16] Xuying Li, "Optimizing Safe and Aligned Language Generation: A Multi-Objective GRPO Approach," arXiv, 2025. [Online]. Available: https://arxiv.org/html/2503.21819v1