International Journal of Computer Applications Technology and Research
Volume 15–Issue 06, 01 – 05, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1506.1001

Reinforcement Learning from Human Feedback (RLHF) and Multimodal LLM Alignment: A Comprehensive Review

Jyotsna Shastry
Department of AIML
Indore Institute of Science and Technology
Indore, India

Dr. Shweta Agrawal
Department of AIML
Indore Institute of Science and Technology
Indore, India

Abstract: Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) with human values and preferences is one of the most important problems in AI research. Reinforcement Learning from Human Feedback (RLHF) is one of the best approaches for addressing these challenges and developing helpful, harmless, and honest AI systems aligned with human values. This paper presents a comprehensive review of RLHF and its multimodal extensions, including their theoretical foundations, algorithmic developments, and practical applications. This survey studies the development of RLHF from foundational reward modeling approaches to advanced methods such as PPO, DPO, RLHF-V, MM-RLHF, and FACT-RLHF. Safe RLHF and Safe RLHF-V are important safety-focused variants, and special attention is also given to red-teaming and robustness evaluation techniques. We also discuss open challenges such as reward hacking, scalable oversight, and integrating factual accuracy into preference learning. The main objective of this survey is to provide a conceptual map and technical reference for researchers and practitioners working on the alignment of foundation models.
Keywords: Reinforcement Learning from Human Feedback, Large Language Models, Multimodal Alignment, Direct Preference Optimization, Safe RLHF, Reward Modeling, AI Safety.

1. INTRODUCTION

The growth of LLMs has developed unprecedented abilities in natural language understanding, text generation, and reasoning. However, raw capability alone is not sufficient for real-world deployment; models must behave in alignment with human intent by being helpful, harmless, and reliable.Model alignment.

Reinforcement Learning from Human Feedback (RLHF) was first proposed for deep reinforcement learning agents.In this approach, the agent learns through human preference signals . After that, the language model was adapted using a fine-tuning strategy [1,2,3] . The main core idea of RLHF is to use reward models trained on human comparative judgments instead of manually designed reward functions.

This enables models to learn nuanced and context-sensitive behaviors that are difficult to define formally [4, 5, 6]. The success of RLHF in text-only systems motivated its extension to multimodal settings, where models must align their behavior across both vision and language modalities. Similarly, safety-aware RLHF variants were developed because deployed AI systems operate in high-stakes environments, requiring the training objective to explicitly incorporate harm avoidance [7, 8].

This review is organized as follows. Section 2 presents the theoretical foundations of RLHF and RL policy optimization. Section 3 surveys RLHF for text-based LLMs. Section 4 covers multimodal alignment methods. Section 5 addresses safety in RLHF. Section 6 examines alternative and complementary alignment strategies. Section 7 discusses open challenges, and Section 8 concludes.

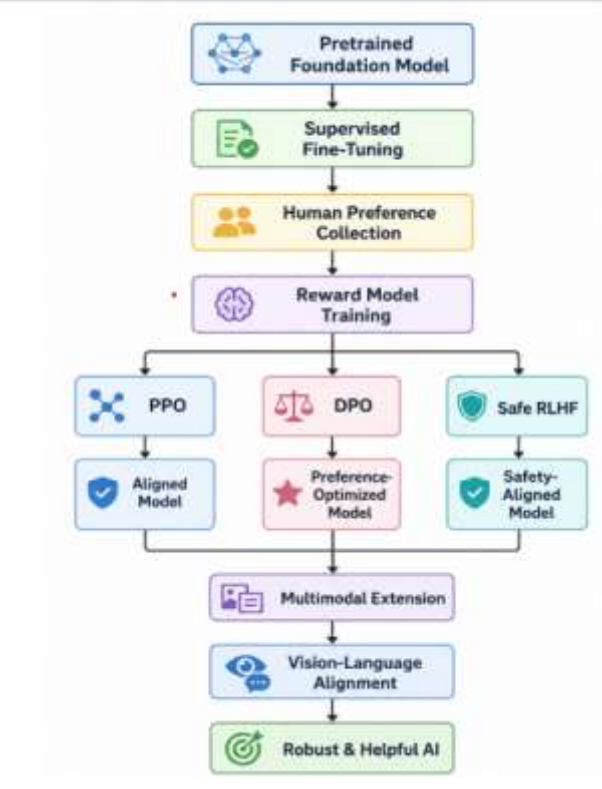

Fig 1. End-to-end survey architecture

International Journal of Computer Applications Technology and Research
Volume 15–Issue 06, 01 – 05, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1506.1001

2. THEORETICAL FOUNDATIONS

2.1 Reinforcement Learning Basics

RLHF is based on the reinforcement learning (RL) framework, where an agent interacts with an environment to maximize cumulative reward. The standard formulation is based on a Markov Decision Process (MDP), which includes states, actions, transition dynamics, and a reward function. The main goal of the agent is to find a policy that maximizes the expected discounted return [9].

2.2 Policy Optimization

One of the most important challenges in RLHF is updating the language model policy in a stable and sample-efficient manner. Proximal Policy Optimization (PPO) solves this problem by constraining the magnitude of policy updates at every training step using a clipped surrogate objective. Due to its stability and empirical effectiveness across different tasks, PPO has become the de facto standard optimizer in RLHF pipelines [2, 3].Theoretical analyses of policy optimization include sample complexity bounds and exploration-exploitation trade-offs [9].

2.3 Reward Modeling from Preferences

The preference-based reward model is a central component of RLHF. Human annotators provide ranked pairs of model outputs, and a Bradley–Terry model is typically used to produce scalar reward scores. These scores are then used to supervise the RL fine-tuning loop. Importantly, the quality and calibration of the reward model directly affect the quality of the aligned policy, making reward model training a major research problem in RLHF [1, 2, 3].

3. RLHF FOR Text-BASED LARGE LANGUAGE MODELS

3.1 Foundational Work

In RLHF, the core loop begins with the policy model generating outputs. Human raters then compare these outputs, and the reward model is trained using these comparisons. After that, the policy is updated through reinforcement learning to maximize the predicted reward. This framework was initially applied to Atari games and simulated robotics.It extended RLHF to abstractive summarization. They showed that fine-tuned models trained with human preference feedback achieved better performance on the TL;DR summarization benchmark compared to supervised baselines. This work established RLHF as a viable technique for large-scale NLP applications.

Fig. 2. RLHF pipeline

3.2 InstructGPT and the Helpfulness/Harmlessness Paradigm

Ouyang et al. [3] introduced InstructGPT, which used large-scale RLHF to train GPT-3 to follow natural language instructions. The three-stage pipeline — supervised fine-tuning (SFT), reward model training on ranked outputs, and PPO-based RL fine-tuning — became the standard template for later systems. Human evaluators consistently found InstructGPT outputs better compared to those of GPT-3. This proves that alignment techniques improve both helpfulness and safety.

Constitutional AI and the 'Helpful and Harmless' (HH) framework were introduced by Bai et al. [1] from Anthropic. In this framework, the reward model was trained not only on helpfulness but also on harmlessness preferences. These harmlessness preferences were obtained from human raters. These preferences were obtained from human raters. This dual-objective reward modeling addresses the tension between usefulness and risk mitigation.

3.3 Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) is one of the most important simplifications of the RLHF pipeline [10]. DPO shows that the RLHF objective can be expressed in closed form as a binary cross-entropy loss over preference pairs As a result, the need for a separate reward model and RL training loop is eliminated. The language model itself implicitly represents the reward function through relative log-probability ratios with respect to a reference model.

DPO has attracted significant attention because of its simplicity and strong empirical performance.The IPO (Identity Preference Optimization) method [11] extended this approach by reinterpreting the preference model as a general classifier. This provides additional theoretical grounding and flexibility.

4. MULTIMODAL LLM ALIGNMENT

4.1 Challenges in Multimodal Settings

Multimodal Large Language Models (MLLMs) process both visual and textual inputs, making their alignment challenges more difficult compared to text-only systems. MLLMs are particularly affected by hallucination problems, where models generate plausible but factually incorrect statements about visual content. Aligning these models requires preference feedback that covers both modalities: vision and language

4.2 Factually Augmented RLHF

Factually Augmented RLHF (Fact-RLHF) was introduced by Sun et al. [4], which integrates factual grounding into the standard preference learning pipeline. The key idea of this approach is that reward models trained only on human

aesthetic preferences may sometimes reward fluent but hallucinated responses. Therefore, Fact-RLHF explicitly incorporates factual accuracy labels into the reward signal.

4.3 RLHF-V: Fine-Grained Correctional Feedback

Yu et al. [6] introduced RLHF-V, which collects fine-grained correctional human feedback at the segment level instead of the response level.Human annotators identify the specific spans in model outputs that contain errors and provide corrections. This creates a dense supervision signal that focuses learning on the exact points where failures occur

Table 1. Multimodal RLHF Methods

Framework	Modalities	Key Contribution	Limitation
RLHF-V	Vision + Language	Fine-grained visual feedback	Expensive annotation
MM-RLHF	Multimodal alignment	Joint reward learning	Large computational cost
Fact-RLHF	Vision + Text	Factuality-aware rewards	Complex verification
Safe RLHF-V	Vision + Text	Safety-oriented alignment	Limited benchmark datasets

5. SAFETY –AWARE RLHF

5.1 Safe RLHF for Text-Based Models

Ji et al. [7] formalized the problem of Safe RLHF. They observed that standard RLHF objectives combine helpfulness and harmlessness into a single scalar reward, making it difficult to explicitly control the trade-off between safety and utility. Safe RLHF separates these objectives by training different reward models for helpfulness and harmlessness. After that, the policy is optimized under a constrained RL formulation, where safety acts as a hard constraint instead of a soft regularizer.

5.2 Red-Teaming and Adversarial Evaluation

This approach is used to generate harmful outputs from the target model. Automated red-teaming scales evaluation beyond human testers and enables testing at large scale. The study showed that automated red-teaming can uncover failure modes that are not detected through manual evaluation. Furthermore, it provides automated toxicity detection tools for evaluating harmful model outputs.

5.3 Safe RLHF-V: Multimodal Safety Alignment

Ji et al. [8] extended the Safe RLHF framework to multimodal models by introducing Safe RLHF-V. This work highlights unique safety challenges present in multimodal systems, where harmful content can be generated in response to visually grounded prompts. For example, images of dangerous activities or weapons combined with instructions may trigger harmful responses. Safe RLHF-V collects multimodal preference data with explicit safety labels and uses constrained policy optimization to penalize unsafe visual-language responses.

5.4 Trustworthiness and the Impact of RLHF on Model Behavior

Li et al. [12] studied the effects of preference alignment on multiple dimensions of trustworthiness, including calibration, robustness, privacy preservation, and fairness. Their study showed that RLHF does not improve trustworthiness in every situation. Although RLHF reduces harmful outputs, it may negatively affect calibration and increase sycophantic behavior. These findings emphasize that RLHF evaluation should not be based only on safety, but also requires a multi-dimensional evaluation framework.

Table 2. Evaluation Metrics for RLHF Models

Metric	Purpose	Example Benchmark
Helpfulness	Measures utility of responses	HumanEval
Harmlessness	Detects unsafe outputs	Toxicity benchmarks
Truthfulness	Evaluates factual correctness	TruthfulQA
Robustness	Measures stability under attacks	Adversarial benchmarks
Alignment Score	Measures human preference agreement	Preference datasets

6. ALTERNATIVE AND COMPLEMENTARY ALIGNMENT METHODS

6.1 Knowledge-Grounded Dialogue

Dinan et al. [13] presented the Wizard of Wikipedia dataset as a benchmark for knowledge-grounded conversational agents. Although this work was conducted before the RLHF era, it established the importance of factual grounding in dialogue systems, which later became an important theme in multimodal RLHF research [5, 7].Knowledge-powered dialogue provides a complementary alignment approach to preference learning. Instead of optimizing models only for human approval, it constrains model outputs to verifiable factual content.

6.2 Model Compression and Distillation for Alignment

MiniLM, a deep self-attention distillation method used to compress pre-trained transformer models. Although MiniLM is not a direct alignment method, it has become important in the RLHF ecosystem through model distillation. Smaller and distilled reward models or policy models help make RLHF computationally feasible at large scale. Furthermore, distillation from aligned teacher models can transfer alignment properties to compact student models without requiring a full RLHF pipeline.

6.3 IPO and Classifier-Based Preference Learning

Garg et al. [13] reformulated preference-based alignment as a classification problem. They showed that language models can function as preference classifiers by using probability distributions over candidate responses. This reinterpretation was formalized as IPO (Identity Preference Optimization), which combines both DPO and standard RLHF as special cases within a unified theoretical framework. Furthermore, this framework supports new algorithmic variants. IPO also achieves strong performance on standard alignment benchmarks with minimal additional training

7. OPEN CHALLENGES AND FUTUTRE DIRECTIONS

7.1 Reward Hacking and Overoptimization

Reward hacking is a fundamental limitation of RLHF. In this problem, the policy learns to take advantage of the weaknesses and idiosyncrasies of the reward model instead of improving the desired behavior. This problem manifests as mode-seeking behavior, excessive verbosity, or sycophantic responses, where the model outputs achieve high scores on reward models but are judged poorly by independent human evaluators. To mitigate reward hacking, better-calibrated reward models, ensemble methods, and KL-divergence penalties are required, as they prevent the policy from deviating excessively from the reference policy [2, 3].

7.2 Scalable Oversight.

As model capabilities increase, it becomes difficult for human evaluators to assess the correctness of complex model outputs. Scalable oversight methods such as debate, recursive reward modeling, and AI-assisted evaluation aim to maintain the fidelity of human feedback even as the complexity of evaluation tasks increases. The automated red-teaming approach [14] is one example in this direction, while the development of multimodal oversight mechanisms for visual reasoning tasks remains an open research frontier [7, 8]. Knowledge-powered dialogue provides a complementary alignment approach to preference learning. Instead of optimizing models only for human approval, it constrains model outputs to verifiable factual content.

7.3 Data Quality and Annotation Consistency

Preference data quality is the main foundation of RLHF. Annotator disagreement, cultural bias, and inconsistent labeling criteria degrade the quality of reward models and introduce unpredictable behaviors in aligned policies. MM-RLHF [7] and the multimodal alignment survey [11] identify data quality as one of the most important bottlenecks in multimodal settings. herefore, rigorous annotation protocols, inter-annotator agreement metrics, and active learning strategies are important for improving preference data collection and remain significant open research problems.

7.4 Multimodal Hallucination

Despite the significant progress of Fact-RLHF [4] and RLHF-V [6], the problem of multimodal hallucination has not yet been fully solved. Current RLHF methods can reduce hallucination, but they cannot completely eliminate it. Similarly, performance degrades on out-of-distribution visual inputs. Integrating visual grounding, object detection feedback, and pixel-level preference signals into RLHF pipelines is considered a promising research direction that requires further investigation.

8. CONCLUSION

This research surveys the full landscape of Reinforcement Learning from Human Feedback (RLHF) and multimodal LLM alignment, discussing the field from foundational principles to the latest developments. RLHF and its variants, such as DPO, Safe RLHF, and multimodal extensions, have established themselves as the dominant paradigm for aligning AI systems with human values and preferences.

The field has evolved from basic reward modeling approaches to advanced multimodal and safety-aware alignment frameworks, while innovations such as DPO have reduced computational complexity and improved accessibility.

There are still many challenges in this field. Reward hacking, hallucination, annotation inconsistency, and the scalability of human oversight are structural challenges that no current method fully solves. As model capabilities continue to increase, multimodal systems are becoming increasingly important in real-world AI deployments. This also increases the need for robust, efficient, and theoretically well-grounded alignment methods. We hope that these methods will help address existing challenges and serve as useful references and roadmaps for future research.

9. REFERENCES

[1] Y. Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," arXiv:2204.05862, 2022.

[2] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in Proc. NeurIPS, vol. 35, 2022, pp. 27730–27744

[3] N. Stiennon et al., "Learning to Summarize from Human Feedback," in Proc. NeurIPS, vol. 33, 2020, pp. 3008–3021.

[4] Z. Sun et al., "Aligning Large Multimodal Models with Factually Augmented RLHF," arXiv:2309.14525, 2023.

[5] X. Yu et al., "RLHF-V: Towards Trustworthy Multimodal Large Language Models via Behavior Alignment from Fine-grained Correctional Feedback," in Proc. CVPR, 2024.

[6] Y. He et al., "MM-RLHF: The Next Step Forward in Multimodal LLM Alignment," arXiv:2502.10391, 2024.

[7] J. Ji et al., "Safe RLHF: Safe Reinforcement Learning from Human Feedback," in Proc. ICLR, 2024.

[8] J. Ji et al., "Safe RLHF-V: Safe Reinforcement Learning from Multimodal Human Feedback," 2025.

[9] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement Learning: Theory and Algorithms," 2022. Available: https://rltheorybook.github.io/

[10] R. Rafailov et al., "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model," in Proc. NeurIPS, 2023.

[11] [11] T. Yu et al., "Aligning Multimodal LLM with Human Preference: A Survey," arXiv.org, 2025.

[12] A. J. Li, S. Krishna, and H. Lakkaraju, "More RLHF, More Trust? On The Impact of Preference Alignment On Trustworthiness," ICLR, 2024.

[13] E. Dinan et al., "Wizard of Wikipedia: Knowledge-Powered Conversational Agents," in Proc. ICLR, 2019.

[14] [S. Garg et al., "IPO: Your Language Model is Secretly a Preference Classifier," ACL, 2025.

[15] E. Perez et al., "Red Teaming Language Models with Language Models," arXiv:2202.03286, 2022.

[16] J. Ji et al., "Safe RLHF-V: Safe Reinforcement Learning from Multi-modal Human Feedback," 2025