RLHF and Alignment

Reinforcement Learning from Human Feedback and Preference Optimization

February 2026 • Technical Report

Table of Contents

1. Introduction

Reinforcement learning from human feedback (RLHF) has become one of the defining techniques in modern large language model development. While pre-training on large text corpora gives models broad knowledge and linguistic competence, it does not teach them to produce outputs that are helpful, harmless, and honest. RLHF and related alignment techniques bridge this gap by incorporating human preferences directly into the model optimization process, steering model behavior toward outputs that humans judge to be high quality.

This report examines the RLHF pipeline, alternative preference optimization methods such as DPO, reward modeling, and the broader landscape of alignment techniques used to make LLMs safe and useful.

2. The Alignment Problem

A pre-trained language model is optimized to predict the next token in a sequence, which is not equivalent to producing responses that are helpful or safe. Pre-trained models will readily generate toxic content, fabricate information, follow harmful instructions, or produce verbose and unhelpful responses because these patterns exist in their training data. The alignment problem is the challenge of modifying model behavior to be consistent with human values and intentions while preserving the model's underlying capabilities. RLHF emerged as the first practical solution to this problem at scale, with InstructGPT demonstrating that a relatively small amount of human feedback could dramatically improve the helpfulness and safety of GPT-3.

3. The RLHF Pipeline

3.1 Supervised Fine-Tuning (SFT)

The RLHF process typically begins with supervised fine-tuning, where the pre-trained model is trained on a curated dataset of high-quality instruction-response pairs. These demonstrations, often written by human annotators, teach the model the basic format of helpful responses: following instructions, providing structured answers, and maintaining appropriate tone. SFT provides a strong starting point but is limited by the diversity and quality of the demonstration data.

3.2 Reward Modeling

A reward model is trained to predict human preferences between model outputs. Annotators are presented with pairs of model responses to the same prompt and asked to indicate which is better. These preference pairs are used to train a separate model (typically initialized from the SFT model) that assigns scalar reward scores to prompt-response pairs. The reward model effectively distills human judgment into a differentiable function that can be used to optimize the language model. Training a

reliable reward model requires careful attention to annotator calibration, inter-annotator agreement, and the diversity of the comparison data.

3.3 Policy Optimization with PPO

The final stage uses reinforcement learning to optimize the language model (the policy) to maximize the reward model's scores. Proximal Policy Optimization (PPO) has been the standard algorithm, adapted from the RL literature. The policy generates responses to a batch of prompts, the reward model scores them, and the policy is updated to increase the probability of high-scoring responses. A KL-divergence penalty against the SFT model prevents the policy from diverging too far, which would lead to reward hacking, where the model learns to exploit quirks in the reward model rather than genuinely improving quality.

PPO-based RLHF is effective but operationally complex. It requires running four models simultaneously (the policy, the reference policy, the reward model, and the value model), making it memory-intensive and sensitive to hyperparameter choices. Training instability, reward hacking, and mode collapse are common failure modes that require careful monitoring and tuning.

4. Direct Preference Optimization (DPO)

DPO was introduced as a simpler alternative to the full PPO-based RLHF pipeline. The key insight is that the reward modeling and RL optimization steps can be collapsed into a single supervised learning objective that directly optimizes the policy on preference data. DPO derives a closed-form expression for the optimal policy given a reward function and uses this to construct a loss function that increases the probability of preferred responses relative to dispreferred responses, with an implicit KL constraint.

DPO eliminates the need for a separate reward model, a value network, and the complex PPO training loop. It requires only the preference dataset and the SFT model, making it far more accessible and stable to train. Since its introduction, DPO and its variants (IPO, KTO, ORPO, SimPO) have been widely adopted, with many organizations using DPO as their primary preference optimization method. However, some research suggests that PPO-based RLHF still outperforms DPO at the frontier of model quality, particularly when the reward model is very strong.

5. Reward Model Design and Challenges

5.1 Bradley-Terry and Beyond

Most reward models are trained using the Bradley-Terry model of pairwise preferences, which assumes that the probability of preferring one response over another is a function of the difference in their underlying reward scores. While this framework is simple and effective, it has limitations: it assumes a single scalar quality dimension, cannot capture context-dependent or multidimensional preferences, and struggles with ties or near-

equal comparisons. Recent work has explored alternatives including Thurstone models, listwise ranking objectives, and multi-objective reward models that separately evaluate helpfulness, harmlessness, and factual accuracy.

5.2 Reward Hacking

Reward hacking occurs when the policy learns to exploit patterns in the reward model that do not correspond to genuine quality improvements. Common manifestations include generating excessively verbose responses (which reward models often rate higher), adopting sycophantic tones that agree with user premises regardless of accuracy, or producing outputs with superficial markers of quality (hedging language, structured formatting) without substantive content. Mitigations include stronger KL penalties, reward model ensembles, iterative reward model retraining, and constitutional AI approaches that use rules-based evaluation alongside learned rewards.

6. Alternative and Complementary Approaches

6.1 Constitutional AI (CAI)

Constitutional AI, developed by Anthropic, reduces reliance on human preference labels by having the model critique and revise its own outputs according to a set of principles (a constitution). In the critique phase, the model identifies problems with its responses based on the principles. In the revision phase, it generates improved responses. These self-generated preference pairs can then be used for RLHF or DPO training. CAI enables scaling of alignment data generation beyond what human annotation alone can support and makes the alignment criteria more explicit and auditable.

6.2 Reinforcement Learning from AI Feedback (RLAIF)

RLAIF extends the CAI concept by using a separate AI model (often a more capable or specially trained model) to generate preference judgments in place of human annotators. Research has shown that RLAIF can produce alignment quality comparable to human-labeled RLHF for many tasks, particularly when the AI judge is well-calibrated and the evaluation criteria are clearly specified. RLAIF dramatically reduces the cost and latency of generating preference data, enabling more frequent alignment iterations during model development.

6.3 Process Reward Models

Standard reward models assign a single score to a complete response (outcome-based reward). Process reward models instead provide feedback on individual reasoning steps, rewarding correct intermediate steps even if the final answer is wrong. This fine-grained supervision is particularly valuable for mathematical reasoning and complex problem-solving tasks, where the model needs to learn not just what the right answer is but how to arrive at it through valid reasoning chains.

7. The Alignment Landscape

Method	Key Mechanism	Complexity	Data Requirements
SFT	Imitation of demonstrations	Low	Curated examples
RLHF (PPO)	RL with reward model	High	Preference pairs + RL infra
DPO	Direct preference optimization	Medium	Preference pairs only
Constitutional AI	Self-critique + revision	Medium	Principles + self-generated data
RLAIF	AI-generated preferences	Medium	AI judge + prompts
Process RM	Step-level reward	High	Step-annotated solutions

8. Evaluation of Alignment

Measuring alignment quality is itself a significant challenge. Automated benchmarks like MT-Bench and AlpacaEval use strong LLMs as judges to evaluate model responses on helpfulness and instruction-following. Safety evaluations test models against taxonomies of harmful content categories using red-teaming prompts and automated classifiers. However, automated metrics can be gamed and may not capture subtle quality differences that humans perceive. Human evaluation remains the gold standard but is expensive, slow, and subject to annotator biases. The most robust evaluation strategies combine automated benchmarks, human evaluation, and targeted red-teaming across multiple dimensions of quality and safety.

9. Future Directions

The alignment field is evolving rapidly. Scalable oversight techniques, where models assist humans in evaluating outputs that are too complex for direct human assessment, are becoming critical as model capabilities increase. Multi-turn and agentic alignment addresses the challenge of steering model behavior across extended interactions and tool-use scenarios, where the consequences of individual actions compound. Mechanistic interpretability research aims to understand how alignment training actually changes model internals, moving beyond behavioral evaluation toward a deeper understanding of why aligned models behave as they do. Finally, the development of robust evaluation frameworks that keep pace with model capabilities remains one of the field's most pressing challenges.

10. Conclusion

RLHF and its descendant techniques have transformed language models from powerful but uncontrolled text generators into useful, safe, and steerable AI assistants. The progression from the complex PPO-based pipeline to more accessible methods like DPO has democratized alignment, while approaches like Constitutional AI and RLAIF have reduced dependence on expensive human annotation. Despite significant progress, the fundamental challenges of reward hacking, scalable oversight, and evaluation robustness ensure that alignment research will remain a central focus of LLM development for the foreseeable future.