

Human Versus AI Feedback for Aligning Large Language Models: A Comparative Review of RLHF and RLAIF

Prof. S. A. Bachwani

Research Scholar, Department of Computer Science and Engineering
Sant Gadage Baba University Amravati, JDIET
Yavatmal, India

Shreya S. Wankhade

Department of Computer Science and Engineering
Jawaharlal Darda Institute of Engineering and Technology
Yavatmal, India

Hema P. Pohankar

Department of Computer Science and Engineering
Jawaharlal Darda Institute of Engineering and Technology
Yavatmal, India

Dr. D. N. Chaudhari

Professor & Dean, Head of Department of Computer Science and Engineering
Jawaharlal Darda Institute of Engineering and Technology
Yavatmal, India

Yash M. Nikhade

Department of Computer Science and Engineering
Jawaharlal Darda Institute of Engineering and Technology
Yavatmal, India

Abstract -Aligning large language models (LLMs) with human preferences commonly relies on reinforcement learning from human feedback (RLHF), which depends on costly preference labels collected from human annotators. Reinforcement learning from AI feedback (RLAIF) replaces some or all of these labels with judgments produced by an LLM. This paper presents a focused comparative review of RLHF and RLAIF based on twelve core publications. We describe the standard RLHF pipeline, organize related AI-feedback methods (Constitutional AI, RLAIF, direct RLAIF, direct preference optimization, self-rewarding models, and generative reward models) in a common taxonomy, and compile the empirical evidence reported in the literature on performance, cost, and the reliability of AI evaluators. In one controlled study, RLAIF matched RLHF on summarization and helpful dialogue and exceeded it on harmless dialogue, whereas other work indicates that the gains from AI feedback depend on the base model, the critic, and how they connect teacher and critic models. Studies of LLM judges further report high agreement with humans alongside position, verbosity, and self-enhancement biases. We conclude that the available evidence supports RLAIF as a way to reduce, but not eliminate, the need for human feedback, and we identify the scarcity of direct head-to-head comparisons as the main open gap. No new experiments were conducted.

Index Terms - Large language models, reinforcement learning from human feedback, reinforcement learning from AI feedback, model alignment, LLM-as-a-judge

I. INTRODUCTION

These pre-trained large language models are trained to predict the next token, which does not guarantee that their

outputs are helpful, accurate, or safe for users. Alignment methods therefore add further training stages. Reinforcement Learning from Human Feedback (RLHF) has emerged as a widely adopted methodology human annotators compare candidate responses; the model learns from these preferences, and its responses is then optimized against that reward [1], [2]. The optimization step typically uses policy-gradient methods such as proximal policy optimization (PPO) [3], and public preference datasets such as the Anthropic helpful and harmless data [4] support this workflow.

A practical obstacle is the dependence on high-quality human preference labels, which are expensive to collect. Reinforcement learning from AI feedback (RLAIF) addresses this by using an LLM to produce the preference labels. The idea was first explored in Constitutional AI [5], and a later study compared human and AI feedback directly on three tasks [6]. Subsequent work questioned how much of the reported benefit of AI feedback is attributable to the feedback itself rather than to the data used for supervised fine-tuning [7], and a separate line of work examined how well LLM judges agree with humans and where they fail [8].

Related methods shorten or reshape the pipeline: direct preference optimization (DPO) removes the explicit reward model [9]; self-rewarding models let the model judge its own outputs [10], and generative reward models use an LLM to produce the reward signal [11]. Application studies have also begun to compare human- and AI-feedback variants in specific domains such as programming feedback [12].

This paper reviews this body of work with three research questions:

- RQ1: How does the performance of RLAIF compare with RLHF in the reported evidence?
- RQ2: What evidence exists on the cost and scalability advantages of AI feedback?
- RQ3: What reliability limits of AI evaluators are reported?

Scope and contributions. This is a literature review. We did not run experiments, train models, or propose a new method; every empirical figure reported here is taken from the cited studies. The contributions are: (i) a unified description of the RLHF and RLAIF pipelines (Section II); (ii) a taxonomy of AI-feedback alignment methods (Section IV); (iii) a consolidated presentation of the empirical evidence from three studies that directly examine AI feedback or AI judges (Section V); and (iv) a synthesis of reliability issues and an identification of gaps (Sections VI and VII).

The rest of the paper is organized as follows. Section II gives background, Section III describes the review scope, Section IV presents the taxonomy, Section V reports the empirical evidence, Section VI discusses reliability, Section VII discusses gaps, limitations and directions, and Section VIII concludes.

II. BACKGROUND: THE RLHF PIPELINE

RLHF is commonly described in three stages [2], [1], [6].

Supervised fine-tuning (SFT). A pre-trained model is fine-tuned on high-quality prompt–response examples to produce an initial policy π^{SFT} .

Reward modelling. For a prompt x , two responses are sampled, and a labeller indicates the preferred response y_w over the non-preferred response y_l . A reward model r_ϕ is trained on such triplets with the Bradley–Terry-style loss [6], [7]:

$$Lr(\phi) = -E(x,y_w,y_l)[\log\sigma(r\phi(x,y_w) - r\phi(x,y_l))], \quad (1)$$

where σ is the sigmoid function.

Reinforcement learning. The policy π_θ , initialized from the SFT model, is optimized to maximize the reward while a Kullback–Leibler (KL) term with coefficient β penalizes drift from the SFT policy [7]:

$$\max E_{x,y \sim \pi_\theta(\cdot|x)} [r_\phi(x,y) - \beta \text{KL}(\pi_\theta(\cdot|x) \parallel \pi^{SFT}(\cdot|x))]. \quad \theta \quad (2)$$

The KL term helps prevent the policy from drifting toward outputs that score highly under the reward model but are of low quality, a failure mode known as reward hacking [6].

The RLAIF variants reviewed below keep this structure but change *who* produces the preference label (y_w,y_l) or the reward, as illustrated in Fig. 1.

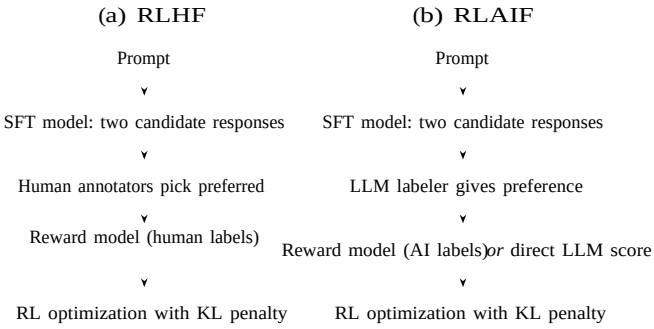


Fig. 1. Schematic comparison of RLHF and RLAIF pipelines. Only the source of the preference signal (shaded) differs. Original schematic drawn by the author from the pipelines described in [6], [1], [2].

Datasets in the reviewed studies. Summarization experiments in [6] use a filtered Reddit TL;DR dataset of about 123k posts and an associated human preference dataset of about 92k comparisons introduced with [2]. Dialogue experiments use the Anthropic helpful and harmless preference data [4].

III. REVIEW SCOPE AND METHOD

This review is *focused*, not systematic. We did not run a database search protocol. The reviewed set of twelve publications was selected because each either (a) introduces or evaluates RLHF or RLAIF, (b) evaluates LLM-as-a-judge reliability, or (c) presents a closely related alternative that also removes or replaces human preference labels. Full-text evidence was extracted from three studies that directly examine AI feedback or AI judges [6], [7], [8]; the remaining works are described at the level of their abstracts and stated contributions. Because the selection is limited and predates any systematic search, the review may omit relevant work, particularly recent publications (see Section VII-B).

IV. TAXONOMY OF AI-FEEDBACK ALIGNMENT METHODS

Table I organizes the reviewed methods by the source of the feedback, the form of the reward, and the optimization step.

TABLE I
TAXONOMY OF THE REVIEWED FEEDBACK-BASED ALIGNMENT METHODS

Method	Feedback source	Reward form	Optimization	Ref.
RLHF	Human annotators	Reward model trained on human preferences	Reinforcement learning with KL regularization (PPO commonly used)	[1], [2]
Constitutional AI	AI model guided by written principles (plus a supervised critique-and-revision phase)	AI-generated harmlessness preferences	Supervised phase, then RL from AI feedback	[5]

Method	Feedback source	Reward form	Optimization	Ref.
RLAIF (canonical)	Off-the-shelf LLM labeler	Reward model trained on soft AI preference labels	REINFORCE with a baseline in [6]	[6]
Direct RLAIF (dRLAIF)	Off-the-shelf LLM scoring outputs during RL	LLM score (1–10, normalized) used directly; no reward model	Same RL setup as canonical RLAIF	[6]
DPO	Preference pairs (human or AI)	Implicit; no separate reward model	Classification-style loss on preference pairs	[9], [7]
Self-rewarding LMs	The model itself as judge (LLM-as-a-judge prompting)	Self-assigned rewards	Iterative DPO	[10]
Generative reward models	LLM used as the reward model	Generated by an LLM	Used within a preference-learning pipeline	[11]

A. Constitutional AI

Constitutional AI [5] trains a harmless assistant through self-improvement without human labels identifying harmful outputs. It has a supervised learning phase followed by a reinforcement learning phase in which the reward signals come from AI feedback. The authors of [6] note that [5] combined LLM-labeled and human-labeled preferences and did not directly compare human and AI feedback, which left open whether RLAIF is a suitable alternative to RLHF.

B. RLAIF and Direct RLAIF

In canonical RLAIF [6], an off-the-shelf LLM is prompted with a preamble, optional few-shot examples, the input, and two candidate responses; the log-probabilities of the tokens “1” and “2” are converted with a softmax into a soft preference (for example 0.6/0.4). To mitigate position bias, each pair is labeled twice with the order reversed and the results are averaged. The reward model is then trained on these soft labels with a cross-entropy loss, which the authors note can be viewed as a form of distillation.

Direct RLAIF removes the reward model: the LLM is prompted to rate a generation from 1 to 10, the score-token likelihoods give a weighted score normalized to $[-1,1]$, and this score is the reward. The stated motivation is that a fixed reward model can become “stale” as the policy drifts away from the distribution on which the reward model was trained [6].

C. Direct Preference Optimization and Self-Rewarding Models

DPO [9] reformulates the objective so that the policy is trained directly on preference pairs with a classification-style loss, skipping the separate reward model and the iterative RL step; it is described as simpler to tune and less computationally demanding than earlier RL-based approaches [7]. Self-rewarding language models [10] let the model act as its own

judge through LLM-as-a-judge prompting and train it with iterative DPO, motivated by the concern that reward models learned from human preferences may be limited by human performance and cannot improve during training.

D. Generative Reward Models

Generative reward models [11] use an LLM in the role of the reward model. We include them for completeness; because this review draws on the abstract-level information for this work, a fuller treatment is left to future versions.

V. REPORTED EMPIRICAL EVIDENCE

A. Direct Comparison of RLAIF and RLHF

Lee et al. [6] compared RLAIF and RLHF on summarization, helpful dialogue, and harmless dialogue. The labeller was PaLM 2 Large, and the policies were initialized from PaLM 2 Extra-Small. Human evaluators compared policy outputs; helpful and summarisation results are win rates, and harmless dialogue is a harmless rate. Table II summarizes the reported numbers.

TABLE II

HUMAN-EVALUATION RESULTS REPORTED IN [6]

Comparison (win rate)	Summarization	Helpful dialogue
RLAIF vs. SFT	71%	63%
RLHF vs. SFT	73%	64%
RLAIF vs. RLHF	50%	52%
d-RLAIF vs. SFT	74%	66%
Same-size RLAIF vs. SFT	68%	—
Harmless rate	SFT 64%, RLHF 76%, RLAIF 88%	

For summarization and helpful dialogue, the differences between RLAIF and RLHF were not statistically significant, and head-to-head win rates were not significantly different from 50%; on harmless dialogue RLAIF scored higher than RLHF [6]. After length adjustment on summarization, the reported win rates against SFT were 59% (RLAIF) and 61% (RLHF). RLAIF improved over SFT even when the labeler had the same size as the policy, and d-RLAIF outperformed same-size canonical RLAIF (preferred 60% of the time) [6].

Several further findings from [6] qualify the headline result. Reward models trained on human labels were more accurate on held-out human preferences than reward models trained on AI labels (for example 79.3% versus 74.2% for summarization), although the final RLAIF and RLHF policies were comparable. Adding AI feedback on top of human feedback did not improve over human feedback alone in the setup tested. AI labeler alignment with human preferences rose with labeler size (62.7% for the smallest, 73.8% for the medium, and 78.0% for the largest model), and chain-of-thought reasoning generally improved alignment, whereas detailed preambles and few-shot examples helped only on some tasks. A small qualitative inspection found cases where RLHF hallucinated and cases where RLAIF was less fluent, but a small blind ranking of 70 examples showed no significant difference. The authors estimate AI labeling to be more than ten times cheaper than human labeling (roughly \$0.06 versus \$0.67 per example under stated assumptions), and note that the estimate excludes factors such as annotator training and setup cost.

B. A Critical Evaluation of AI Feedback

Sharma et al. [7] examined whether the RL step with AI feedback is warranted, using DPO on ShareGPT prompts with several open base models, GPT-3.5, GPT-4 and Claude as teachers or critics, and AlpacaEval for evaluation. They report that RLAIIF substantially improves over supervised fine-tuning on the default GPT-3.5-generated completions, but that supervised fine-tuning on completions from the stronger critic model matches or outperforms RLAIIF, and that using the strong model as teacher in the SFT stage leaves little or no benefit from the additional AI-feedback stage. They conclude that two conditions appear necessary for RLAIIF to clearly beat SFT: a sufficiently strong base model and a capability mismatch between the teacher and the critic; gains also varied across base model families, evaluation protocols, and critics. The authors explicitly caution against generalizing these findings to RLHF. Note that this setting (DPO, compared against SFT, judged by the same critic model) differs from that of [6] (reward model and REINFORCE, compared against RLHF, judged by humans), so the two sets of results are not directly interchangeable.

C. Reliability of LLM Judges

Zheng et al. [8] studied LLM-as-a-judge using MT-Bench (80 multi-turn questions, 58 expert labelers) and a sample of 3K Chatbot Arena votes. Table III lists selected findings. GPT4 judgments agreed with human experts 85% of the time on non-tie votes, compared with 81% agreement among humans. The study also documented biases: position bias (consistency after swapping answer order was 65.0% for GPT-4, 46.2% for GPT-3.5, and 23.8% for Claude-v1 under the default prompt), verbosity bias (Claude-v1 and GPT-3.5 failed a “repetitive list” test in 91.3% of cases and GPT-4 in 8.7%), and self enhancement (GPT-4 and Claude-v1 gave themselves higher win rates by about 10% and 25%, although the authors state the data cannot establish this bias). Judges also struggled with math and reasoning questions: a default prompt failed on 14 of 20 math questions, chain-of-thought on 6, and a reference-guided prompt on 3. Mitigations included swapping positions and calling the judge twice, few-shot examples (raising GPT4 consistency to 77.5%), and reference-guided grading. The study focused on helpfulness and largely did not address safety.

TABLE III
 SELECTED LLM-JUDGE FINDINGS REPORTED IN [8]

Finding	Reported value
GPT-4 vs. human agreement (non-tie, MT-Bench)	85%
Human vs. human agreement (non-tie)	81%
Consistency after swapping order (default prompt)	GPT-4 65.0%, GPT-3.5 46.2%, Claude-v1 23.8%
Failure on “repetitive list” test	Claude-v1 91.3%, GPT-3.5 91.3%, GPT-4 8.7%
Math-grading failures (of 20)	Default 14, CoT 6, reference-guided 3

D. Application Evidence

An application study on automated code feedback [12] fine-tunes LLMs with RLHF and reports, in its abstract, accuracy improvements of 2–5% over RL-free state-of-the-art techniques in automated evaluation and similar or slightly higher accuracy compared with reward-model-free RLAIIF. We

cite it only as an example of domain-specific comparison; we did not verify its full experimental details.

VI. COMPARATIVE ANALYSIS AND RELIABILITY CHALLENGES

Table IV compares RLHF and RLAIIF along dimensions for which the reviewed literature offers evidence. Cells that would require unsupported claims are omitted.

TABLE IV

COMPARISON OF RLHF AND RLAIIF BASED ON REPORTED EVIDENCE

Dimension	RLHF	RLAIIF	Ref.
Feedback source	Human annotators	LLM labeler or judge	[6]
Label cost	Higher (human labeling estimated at about \$0.67 per example in one analysis)	About \$0.06 per example in the same analysis; estimate excludes setup and training costs	[6]
Preference-model accuracy on human labels	Higher (e.g., 79.3% on summarization)	Lower (74.2%) but comparable final policies	[6]
End-policy quality	Preferred over SFT (73% and 64%)	Comparable (71% and 63%); higher harmless rate (88% vs. 76%)	[6]
Sensitivity to labeler/critic	Depends on annotator quality (not quantified in reviewed work)	Alignment rises with labeler size; gains depend on base model, critic and teacher–critic gap	[6], [7]
Known evaluator biases	Not analyzed in the reviewed studies	Position, verbosity, self-enhancement, limited math/reasoning grading	[8], [6]
Oversight recommendation	Gold standard in high-stakes domains, per the authors	Use with caution in such domains	[6]

Position bias. It was observed for both large and small labelers, and it was more pronounced for smaller ones in [6] (the same position was preferred after swapping in 18%, 21%, and 56% of cases for the large, medium and small labeler), and in all judges tested in [8]. Order-swapping with averaging [6] or with a consistency requirement [8] are the reported mitigations.

Verbosity and length. Judges can favor longer answers [8], and RL-trained policies in [6] tended to produce longer responses than SFT, which the authors controlled for post hoc while noting the correction is imperfect.

Self-enhancement and circularity. Judges may favor their own outputs [8]. In [7], the critic and the evaluator were the same model to reduce mismatch, which is reasonable for a controlled comparison but means the measured gains reflect that critic’s own preferences.

Capability dependence. AI-labeler alignment depends on labeller capability [6], and benefits over SFT depend on the teacher–critic relation [7].

Bias transfer and misuse. The authors of [6] note that biases in the off-the-shelf LLM may be transferred into the preferences and amplified by the trained policy, and that

lowering the barrier to alignment also lowers the barrier to misuse.

VII. DISCUSSION

A. Synthesis and Research Gap

Taken together, the three principal studies support a cautious reading. First, in the direct comparison of [6], AI feedback produced policies comparable to RLHF on the tasks and model family tested. Second, [7] shows that the benefit of AI feedback over supervised fine-tuning can depend on the data used for SFT, the base model, and the critic, but it compares against SFT, not RLHF. Third, [8] shows that strong LLM judges can approximate human preferences but exhibit measurable biases.

Gap (our reading of the reviewed literature). Direct head-to-head comparisons of RLHF and RLAIF with human evaluation are scarce in the reviewed set: the principal one [6] uses a single model family and three tasks, and the critical study [7] does not include an RLHF arm. It therefore remains unclear how far the parity reported in [6] extends across model families, tasks, and judge models. Resolving this gap requires new controlled experiments, which are outside the scope of this review.

B. Limitations of This Review

The review is not systematic, covers twelve publications, and extracted full-text evidence for three; some methods (self-rewarding models, generative reward models, application studies) are described only at the level of their abstracts. The reviewed set ends in 2025 and does not cover later work. Numerical results are quoted from the source papers and depend on their models, datasets, and evaluation protocols; they should not be compared across studies without care. No independent experiments were performed.

C. Future Directions

Directions suggested by the reviewed work include: combining human and AI feedback in setups other than the one tested in [6], where a simple combination did not help but the authors propose alternatives such as using RLAIF as a warm-up before RLHF; studying reward-model staleness and iterative labeling [6]; adapting judges to safety criteria, which [8] did not evaluate; and extending direct comparisons of human and AI feedback to more model families and domains, as suggested by the gap above.

VIII. CONCLUSION

This review compared RLHF and RLAIF using twelve core publications. The best direct evidence [6] reports that RLAIF matched RLHF on summarization and helpful dialogue and exceeded it on harmless dialogue, at a labeling cost estimated at more than ten times lower, while human-label reward models remained more accurate and simple human-AI combination did not help in the setup tested. Complementary work shows that benefits of AI feedback depend on the base model, the critic and the SFT data [7], and that LLM judges agree well with humans yet exhibit position, verbosity, and self-enhancement biases [8]. RLAIF therefore appears to be a promising way to reduce, not eliminate, dependence on human feedback, and human oversight remains important, especially in high-stakes domains. Further controlled head-to-head studies are needed.

ACKNOWLEDGMENT

The author sincerely thanks **Dr. D. N. Chaudhari, Head of the Department**, for his valuable guidance and support throughout this work. The author also expresses sincere gratitude to **Prof. S. A. Bachwani** for his guidance and valuable support during the preparation of this work.

REFERENCES

- [1] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–27744.
- [2] N. Stiennon *et al.*, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3008–3021.
- [3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.
- [4] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv:2204.05862, 2022.
- [5] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, 2022.
- [6] H. Lee *et al.*, "RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback," in *Proc. 41st Int. Conf. Mach. Learn.*, PMLR 235, 2024, pp. 26874–26901.
- [7] A. Sharma, S. Keh, E. Mitchell, C. Finn, K. Arora, and T. Kollar, "A critical evaluation of AI feedback for aligning large language models," in *Advances in Neural Information Processing Systems*, 2024, pp. 29166–29190.
- [8] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36, Datasets and Benchmarks Track, 2023.
- [9] 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," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 53728–53741.
- [10] W. Yuan *et al.*, "Self-rewarding language models," in *Proc. 41st Int. Conf. Mach. Learn.*, PMLR 235, 2024, pp. 57905–57923.
- [11] D. Mahan *et al.*, "Generative reward models," arXiv:2410.12832, 2024.
- [12] T. Rahman, S. A. P. Kumar, S. Jha, and A. Ramanathan, "ACERLHF: Automated code evaluation and Socratic feedback generation tool using large language models and reinforcement learning with human feedback," arXiv:2504.04657, 2025.