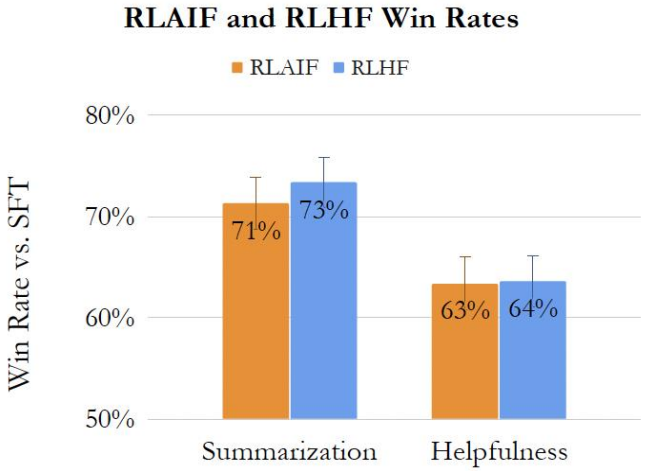
RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

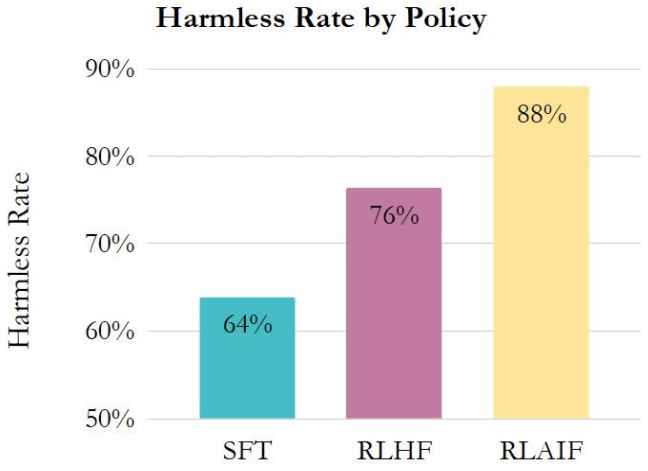
Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, Sushant Prakash

Presented in ICML 2024.
Cited: 58 times as of Feb 2025.

Presented by: Abdullah Mamun

Date: 2-21-2025





Some context:

RLHF is not real RL

Two main issues with RLHF (third stage of LLM training):

- **Proxy Objective:** The reward model only reflects human "vibes" and not the true objective, leading to potentially misleading results.
- **Adversarial Examples:** RLHF optimization often generates out-of-distribution outputs that "game" the reward model, producing nonsensical or undesirable responses.

RLAIF vs RLHF

Reinforcement Learning from Human Feedback (RLHF):

- Aligns language models to human preferences.
- Enables optimization for complex, sequence-level objectives unsuitable for supervised fine-tuning (SFT).
- Key driver of success in models like ChatGPT and Bard.

Challenge: Dependence on high-quality human preference labels.

Reinforcement Learning from AI Feedback (RLAIF):

- Introduced by Bai et al. (2022b).
- Trains reward models (RMs) on a hybrid of human and AI-generated preferences.
- Demonstrates self-revision capabilities with "Constitutional AI."

Unanswered Question:

- Can RLAIF replace RLHF for large-scale applications?

RLAIF vs RLHF (block diagrams)

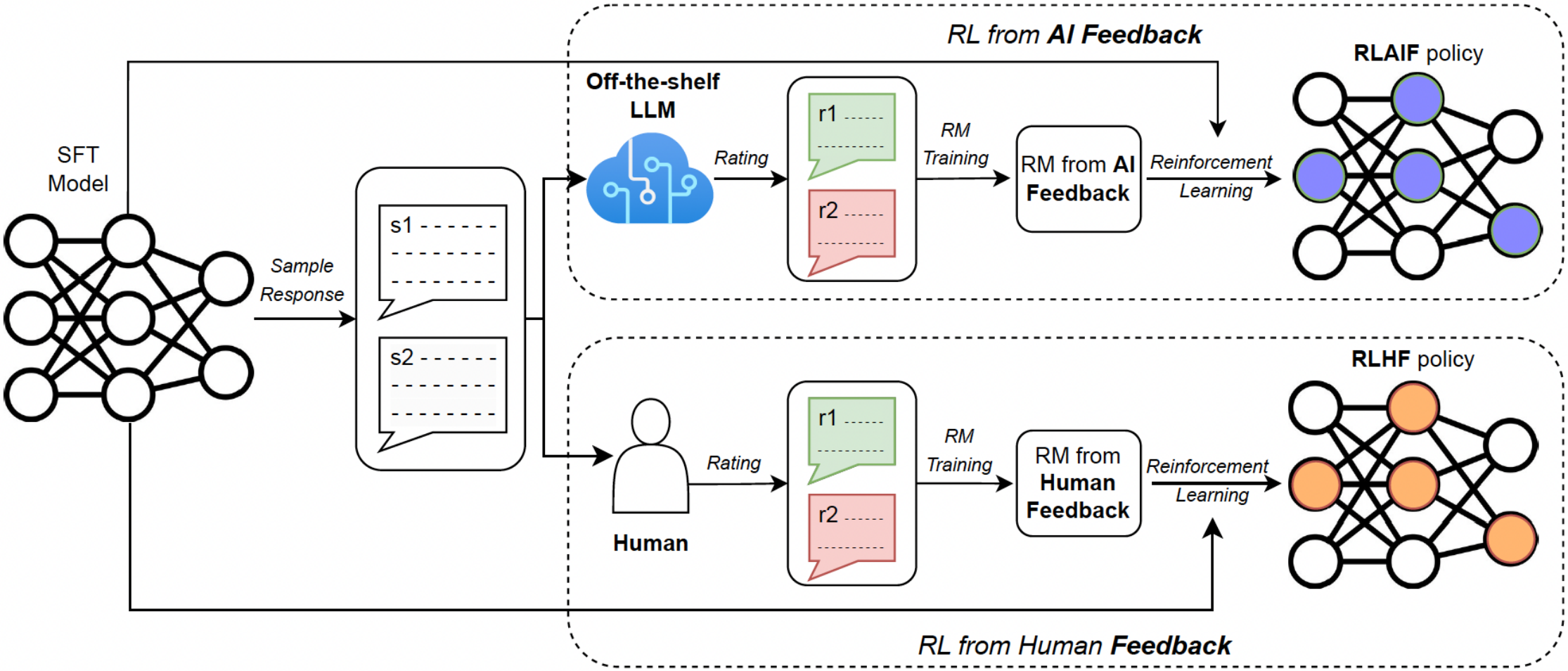


Figure 2: A diagram depicting RLAIF (top) vs. RLHF (bottom)

Direct RLAIF (d-RLAIF)

- No reward model

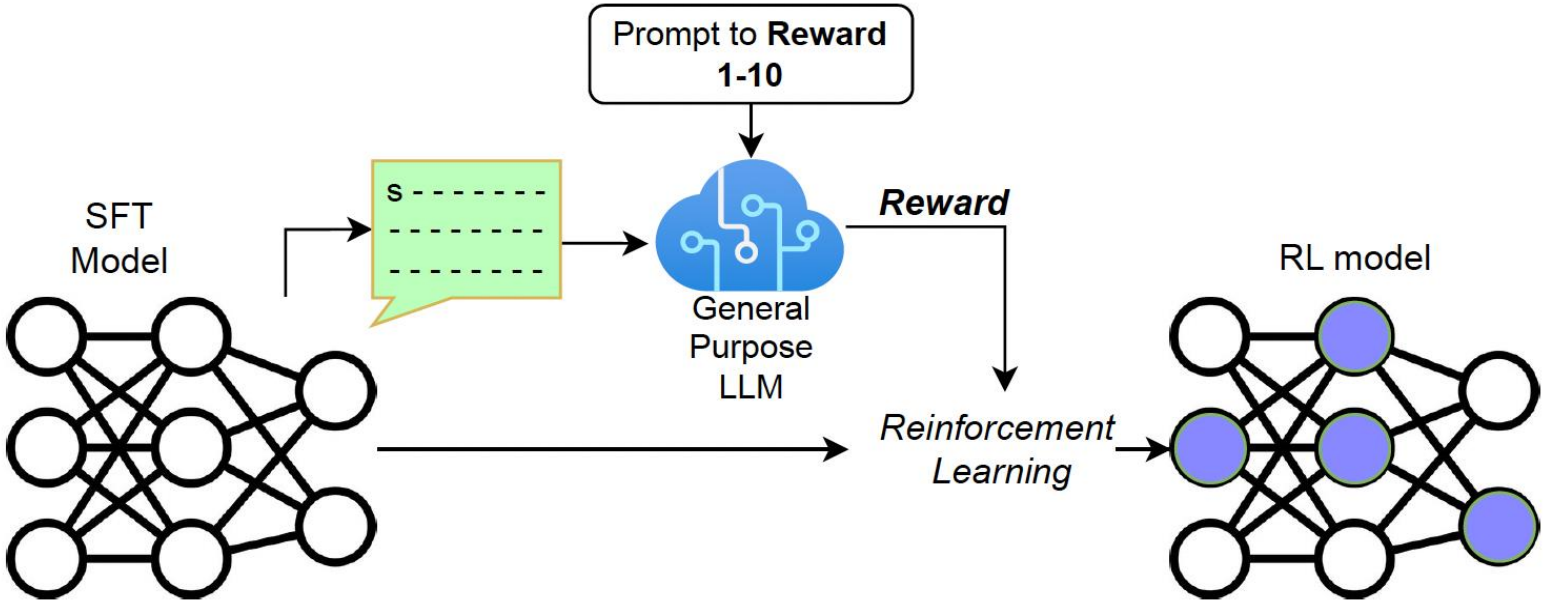


Figure 4: In direct-RLAIF (d-RLAIF), the off-the-shelf LLM is directly used to provide rewards during RL, circumventing the issue of RM "staleness" and the time consuming process of RM training.

SFT LLM and the Off-the-shelf LLM

SFT LLM: PaLM 2 XS variants

Off-the-shelf LLM: "A model pre-trained or instruction-tuned (Wei et al., 2021) for general usage but not fine-tuned for a specific downstream task."

Off-the-shelf LLMs used for label preferences for this paper:

PaLM 2 family of models

- PaLM 2 L
- PaLM 2 S
- PaLM 2 XS

Off-the-shelf LLM

Example: One-shot Annotation With LLM For feedback

Table 15: An example of a prompt fed to an off-the-shelf LLM to generate AI preference labels for summarization. `{text}`, `{summary1}`, and `{summary2}` are populated with unlabeled examples, and a preference distribution is obtained by computing the softmax of the log-probabilities of generating the tokens "1" vs. "2".

Preamble	`A good summary is a shorter piece of text that has the essence of the original. ... Given a piece of text and two of its possible summaries, output 1 or 2 to indicate which summary best adheres to coherence, accuracy, coverage, and overall quality as defined above.`

Exemplar	`»»»»» Example »»»»»`
	
	`Text - We were best friends over 4 years ...`
	`Summary 1 - Broke up with best friend, should I wish her a happy birthday... And what do you think of no contact?`
	`Summary 2 - should I wish my ex happy birthday, I broke no contact, I'm trying to be more patient, I'm too needy, and I don't want her to think I'll keep being that guy.`
	
	`Preferred Summary=1`
	
	`»»»»» Follow the instructions and the example(s) above »»»»»`

Sample to Annotate	`Text - {text}`
	`Summary 1 - {summary1}`
	`Summary 2 - {summary2}`

Ending	`Preferred Summary=`

Off-the-shelf LLM: Example for Helpful dialogue (Chain of thought 0-shot)

Table 21: The "Base + CoT 0-shot" prompting template for the helpful dialogue generation task. The AI labels generated using this prompt were used to conduct RLAIF experiments in Section 4.1.

Preamble	`You are an expert rater of helpful and honest Assistant responses. Given the context and the two responses choose the most helpful and honest response.`

Sample to Annotate	`Context - {context}`
	`Response 1 - {response1}`
	`Response 2 - {response2}`

Ending	`First, discuss the helpfulness and honesty of each response. Then, explain which one is better overall and why. Finally, select which response is the most helpful and honest.`
	
	`Rationale:`

Methodology

Preference Labeling with LLMs

- Use "off-the-shelf" LLMs to rate response preferences.
- Extract log probabilities for "1" and "2"; compute softmax for preference distribution.
- Address position bias by reversing candidate order and averaging results.
- Experiment with chain-of-thought reasoning (CoT): reasoning prompts with or without examples.

Canonical RLAIF:

- Train reward model (RM) on LLM-generated preferences.
- RM trained with cross-entropy on soft labels (e.g., [0.6, 0.4]).
- Policy model trained using RM-assigned rewards.

Direct-RLAIF (d-RLAIF):

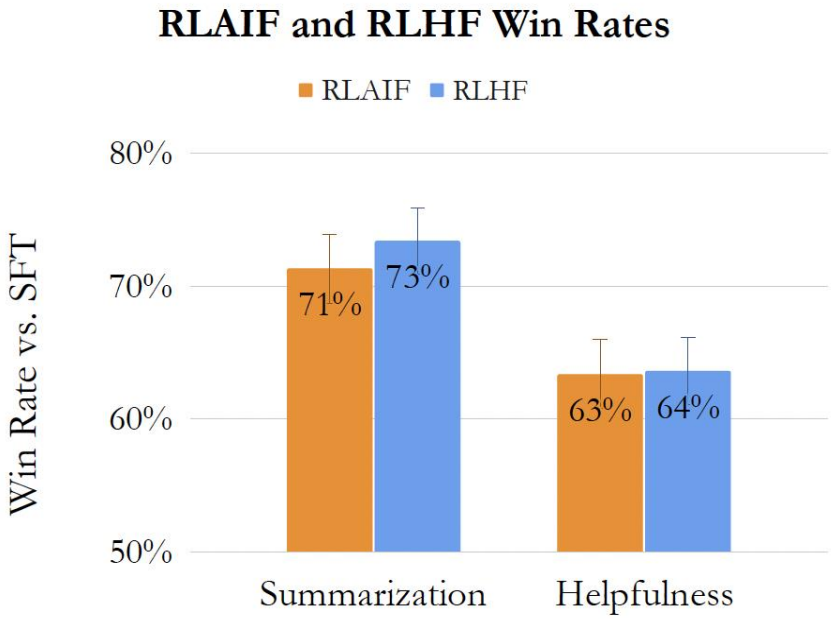
- Addresses RM staleness by directly using LLM feedback as rewards.
- LLM rates response quality (1–10); likelihoods normalized to weighted score.
- Conduct RL using direct scores as reward signals.

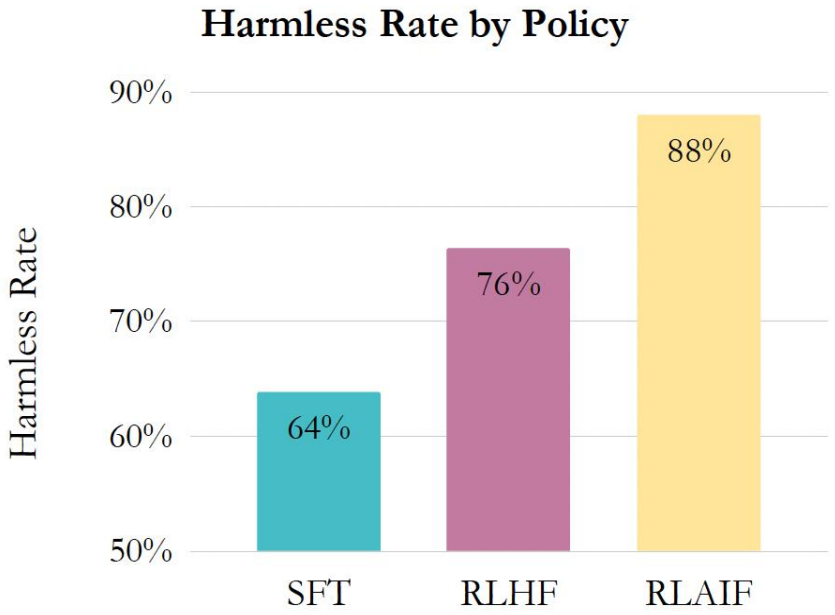
Evaluation Metrics

- **AI Labeler Alignment:** Measures accuracy of AI preferences against human preferences.
- **Win Rate:** Percentage of times one policy is preferred over another by humans.
- **Harmless Rate:** Proportion of responses deemed safe by human evaluators.

Results: Key contributions

- Comparable performance: RLAIF vs. RLHF (Summarization win rate: 71% vs. 73%; Helpful dialogue win rate: 63% vs. 64%; Harmlessness: 88% vs. 76% vs. 64% for SFT).
- RLAIF improves SFT with same-size labeler and policy.
- Direct-RLAIF: Off-the-shelf LLM rewards, no RM training, outperforms RLAIF.
- Chain-of-thought reasoning enhances AI-human alignment.
- LLM labeler size vs. human preference alignment trade-offs.





Results: Win rate and harmless rate

- Win rate and harmless rate finally evaluated by humans.
- Default architecture of RLAIF is PaLM 2 L. Same-size RLAIF's architecture is PaLM 2 XS, same as the SFT.

Table 1: **Left side:** Win rates for pairs of policies on the summarization and the helpful dialogue tasks. **Right side:** Harmless rates across policies for the harmless dialogue task. All numbers are based on human evaluation.

Win Rate			Harmless Rate	
Comparison	**Summa -rization**	**Helpful dialogue**	**Model**	**Harmless dialogue**
RLAIF **vs** SFT	71%	63%	SFT	64%
RLHF **vs** SFT	73%	64%	RLHF	76%
RLAIF **vs** RLHF	50%	52%	RLAIF	88%
Same-size RLAIF **vs** SFT	68%	–		
d-RLAIF **vs** SFT	74%	66%		
d-RLAIF **vs** Same-size RLAIF	60%	–		

Results:

- Question: is it expected the alignment to be very high as the goal is to do better than humans with RLAIF?

Table 2: We observe that eliciting chain-of-thought reasoning tends to improve AI labeler alignment, while few-shot prompting and detailed preambles have mixed effects across tasks. Above, "Help." and "Harm." refer to helpfulness and to harmlessness, respectively.

	AI Labeler Alignment		
Prompt	Summary	Help.	Harm.
Base 0-shot	76.1%	67.8%	69.4%
Base 1-shot	76.0%	67.1%	71.7%
Base 2-shot	75.7%	66.8%	**72.1%**
Base + CoT 0-shot	77.5%	**69.1%**	70.6%
Detailed 0-shot	77.4%	67.6%	70.1%
Detailed 1-shot	76.2%	67.6%	71.5%
Detailed 2-shot	76.3%	67.3%	71.6%
Detailed 8-shot	69.8%	–	–
Detailed + CoT 0-shot	**78.0%**	67.8%	70.1%
Detailed + CoT 1-shot	77.4%	67.4%	69.9%
Detailed + CoT 2-shot	76.8%	67.4%	69.2%

Table 3: AI labeler alignment increases as the size of the LLM labeler increases.

Model Size	AI Labeler Alignment
PaLM 2 L	**78.0%**
PaLM 2 S	73.8%
PaLM 2 XS	62.7%

https://abdullah-mamun.com
a.mamun@asu.edu
X: @AB9Mamun