

# **Reinforcement Learning from Human Feedback (RLHF)**

Ese Ben

Transformers Reading Group

02 October 2023

# RLHF: summary

RLHF: RL + HF

## RL

machine learning for interactive agents

## HF

Human feedback on model's response to learn a **reward model**

# RL... in a nutshell

## Environment

Markov Decision Process (MDP).

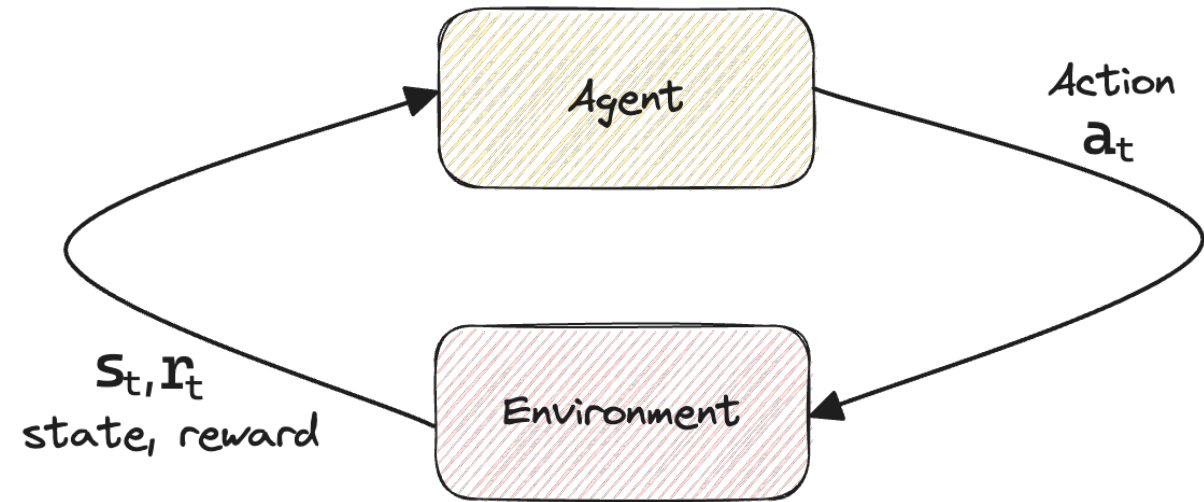
Discrete timesteps  $t$ .

## Agent

Defined as a Policy.

## Active learning

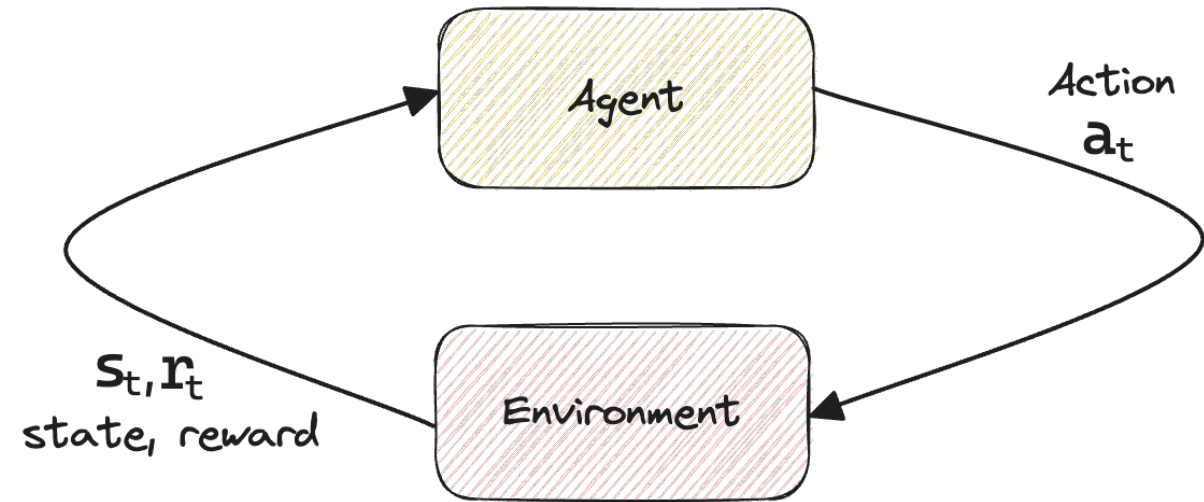
**agent** interacts with **environment** and receives **reward**.



# RL... in a nutshell

**Environment (MDP):** a tuple  
modelled as a tuple:  $\langle S, A, P, R \rangle$

- $S$ : set of states
- $A$ : set of actions
- $P$ :  $p(s_{t+1}|s_t, a_t)$  transition function
- $R$ :  $r(s_t, a_t)$  reward function



# RL... in a nutshell

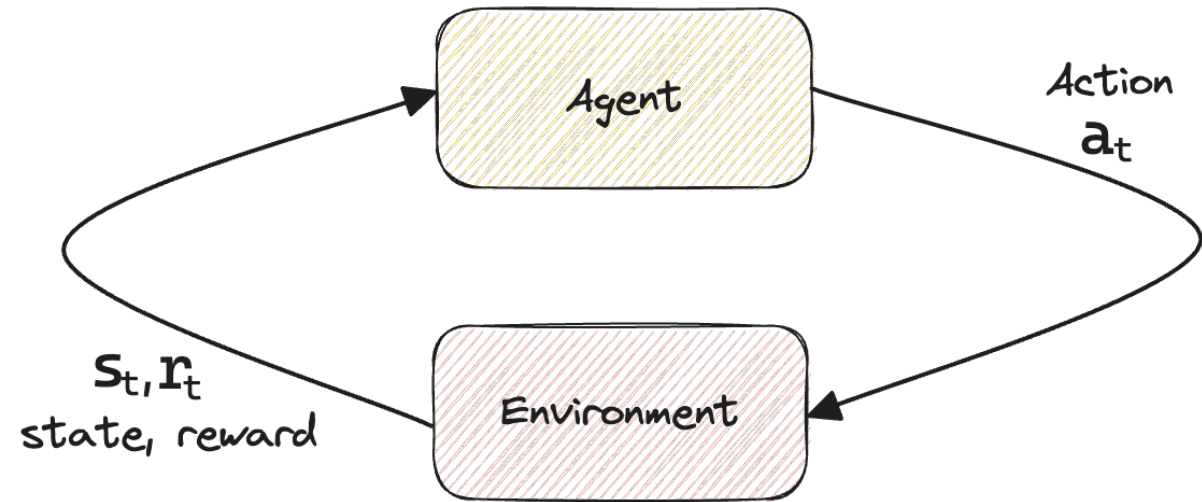
## Agent

defined as a Policy:  
stochastic  $\pi$  or deterministic  $\mu$ .

## Goal

maximise reward obtained.

$$\sum_{t=0}^n r_t$$



# RL... in a nutshell

## Procedure: General RL steps (Algo)

Initialise policy

for each train step

- obtain data (state, action, reward) following policy.

- store data in buffer.

- optimise policy using data.

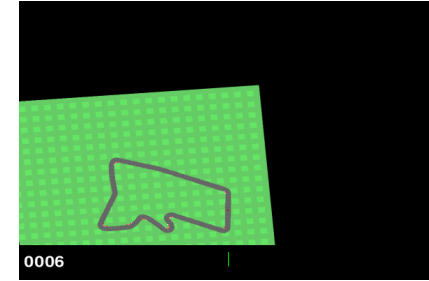
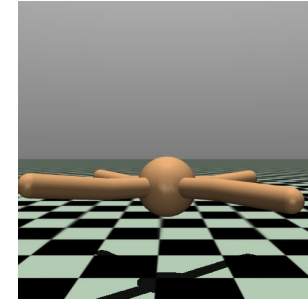
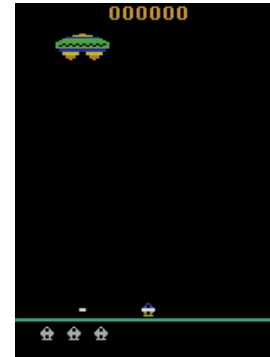
# RL... in a nutshell

## Use Cases

Popularly used in robotics and games.

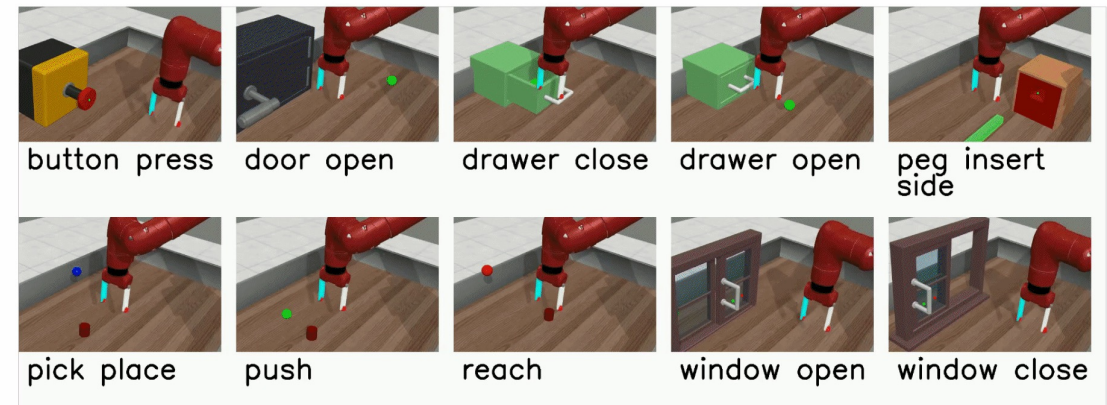
## Now

Language modelling



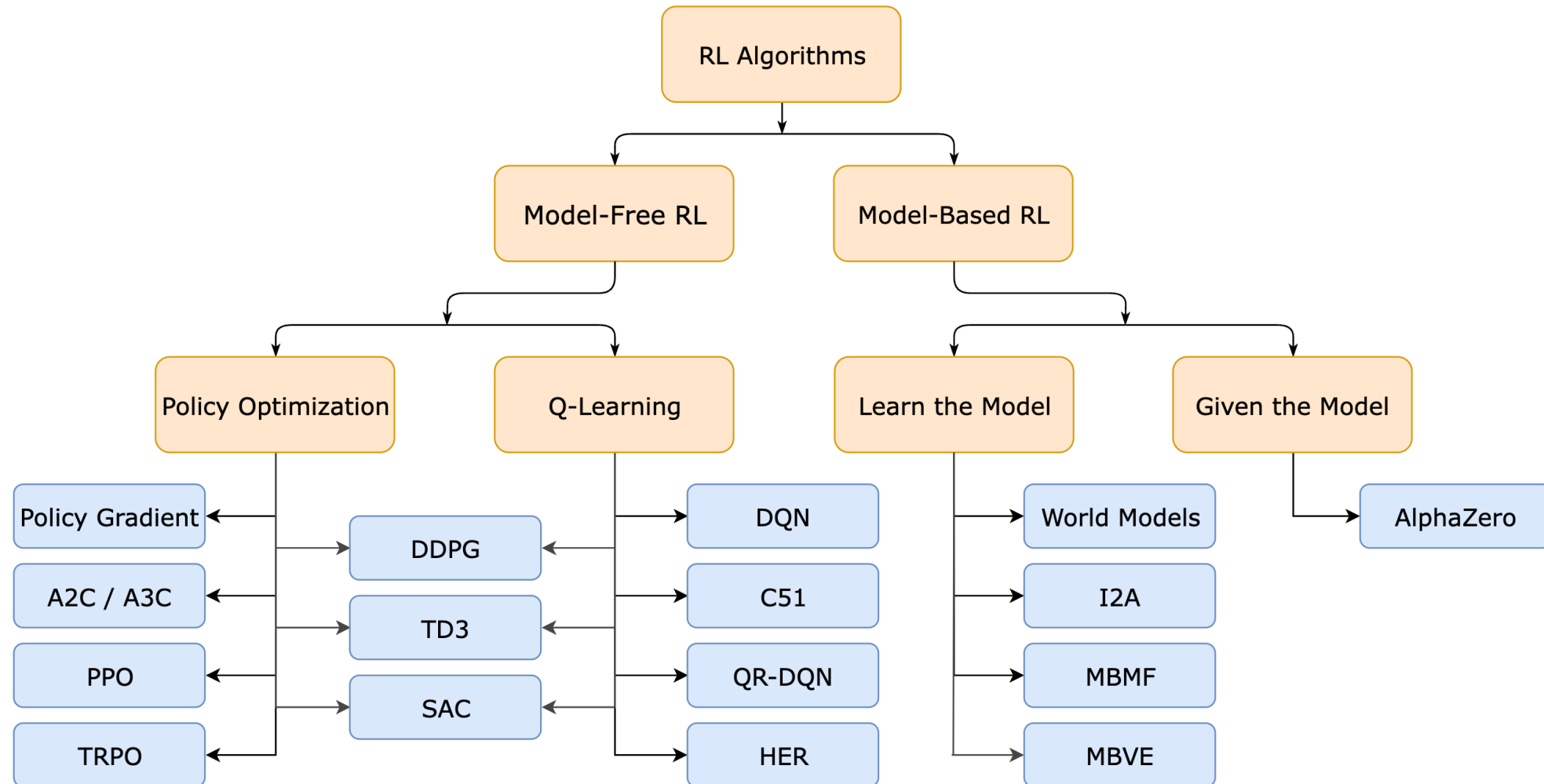
Source: <https://gymnasium.farama.org/>

Train



Source: <https://meta-world.github.io/>

# RL Taxonomy



Source: [https://spinningup.openai.com/en/latest/spinningup/rl\\_intro2.html](https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html)



**So RLHF...**

# Previously, SFT/PEFT...

- Train base LLM to respond to prompt/instructions/queries.
- Attempts to produce **helpful, honest, harmless (3H)** answers.
- Grounded in human preferences via alignment.

## But...

- Not all responses to same prompt are equal.
- Can we ground model towards the most useful response?
- How to define **complex** objective functions for response that are funny or harmless or safe or ethical...?

**We model it!**

# RLHF

0. Acquire data:

- **Generate** multiple response from each prompt.
- Human operator **evaluate quality** of each response.

1. Learn reward model from data

2. Use reward model to finetune LLM via RL (PPO Algo.  
Commonly used)

# RLHF

## Feedback interface

Source:  
<https://icml.cc/virtual/2023/tutorial/21554>

human rates better response



Hugging Face



Toloka

**Human**  
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

**Assistant**  
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

**Human**  
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

**Assistant**  
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

**Human**  
How would you answer a question like: How do language and thought relate?

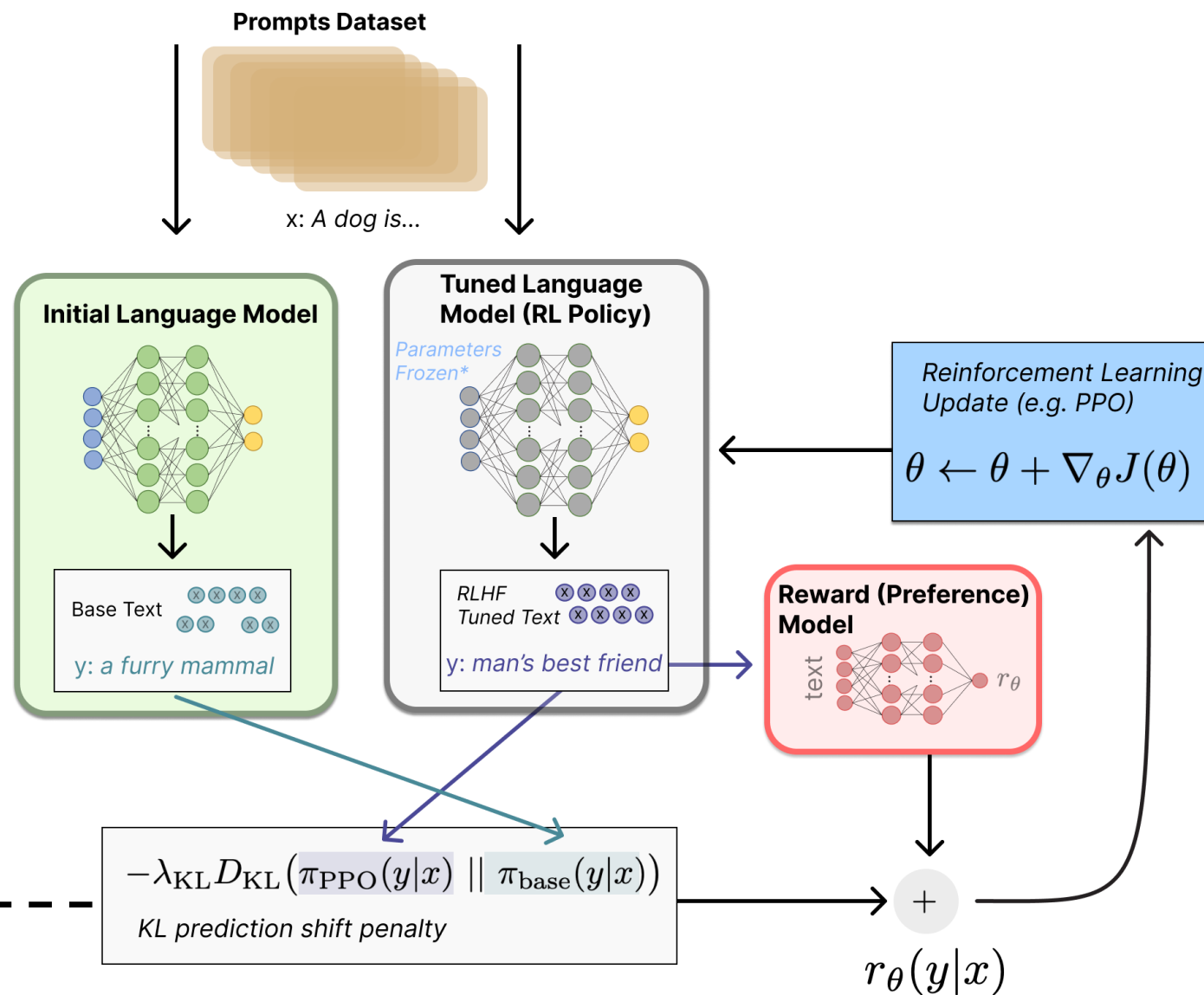
**Choose the most helpful and honest response**

A  
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

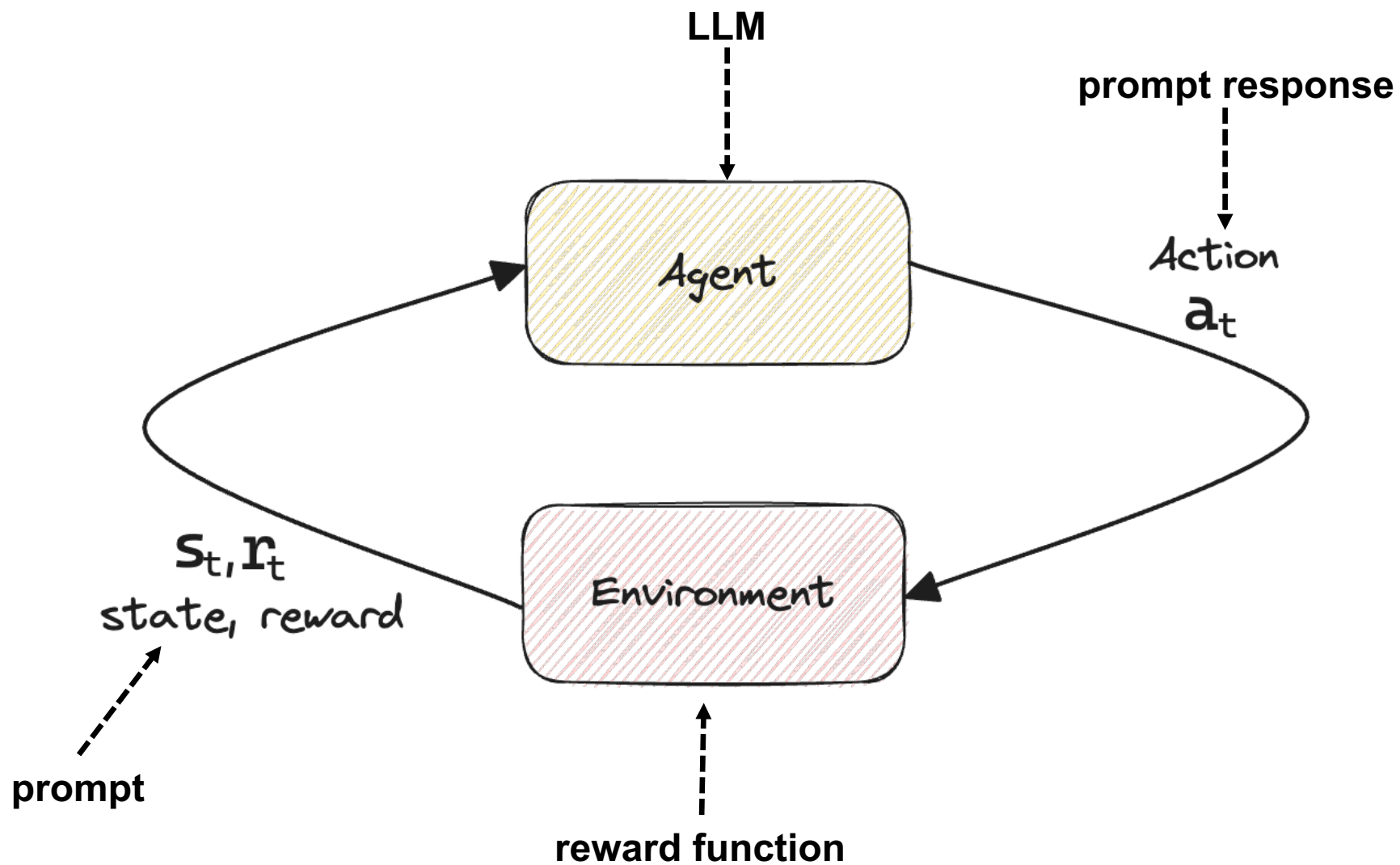
B  
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A is better      B is better

# RLHF



# RLHF in RL Lens

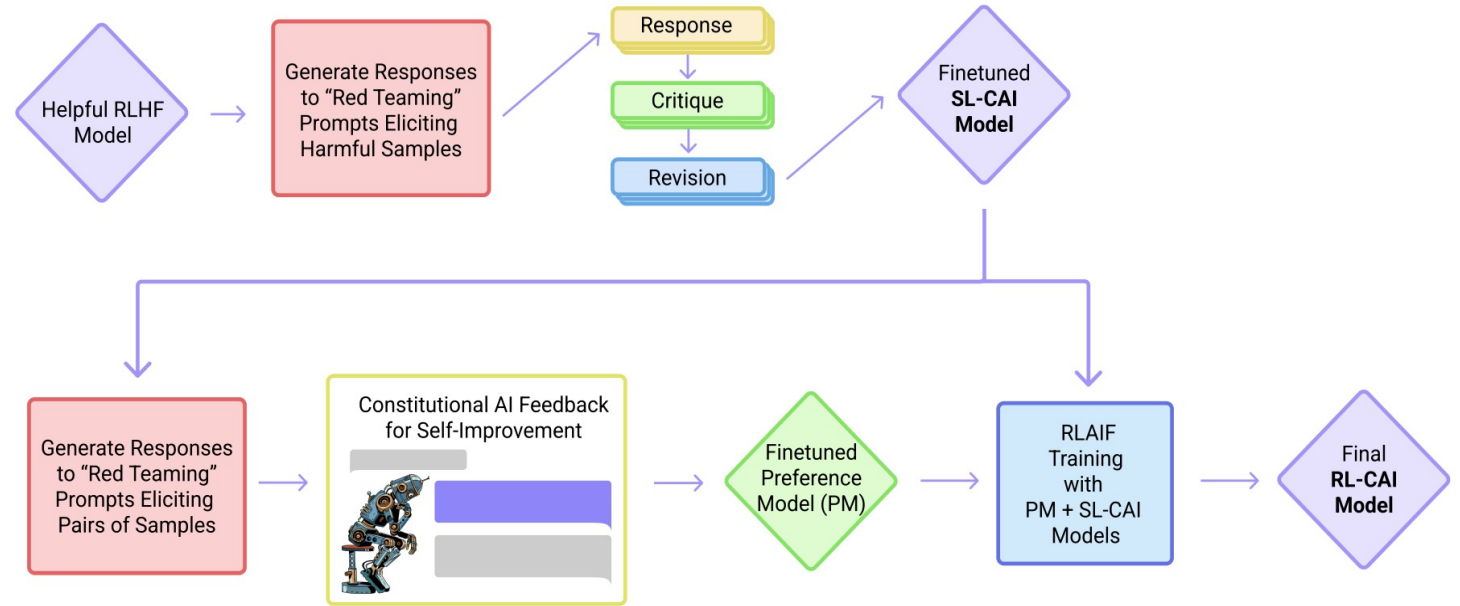


# **Other Approaches**

# RLAIF

## RL from AI Feedback

- Like RLHF, but with very little human feedback.
- Human involvement encoded into a **Constitution** used in prompting the model



**Figure 1** We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Source: <https://arxiv.org/abs/2212.08073>



# ReST

## REinforced Self Training (ReST)

- Follows RLHF philosophy (learn a RM, and then train with RL)
- **Offline RL** instead of online RL.
- Two steps: Improve and Grow

Algorithm	Average Reward	Distinct samples
BC (G=0, I=0)	70.9	16 000 000
<i>ReST</i> (G=1, I=0)	71.9	16 000 000
<i>ReST</i> (G=1, I=4)	77.8	16 000 000
<i>ReST</i> (G=2, I=3)	<b>83.1</b>	32 000 000
Online RL	71.6	24 000 000

Table 1 | **Online RL for IWSLT 2014:** Online RL performs as well as *ReST* (G=1, I=0) and *ReST* (G=1, I=4) is significantly better.

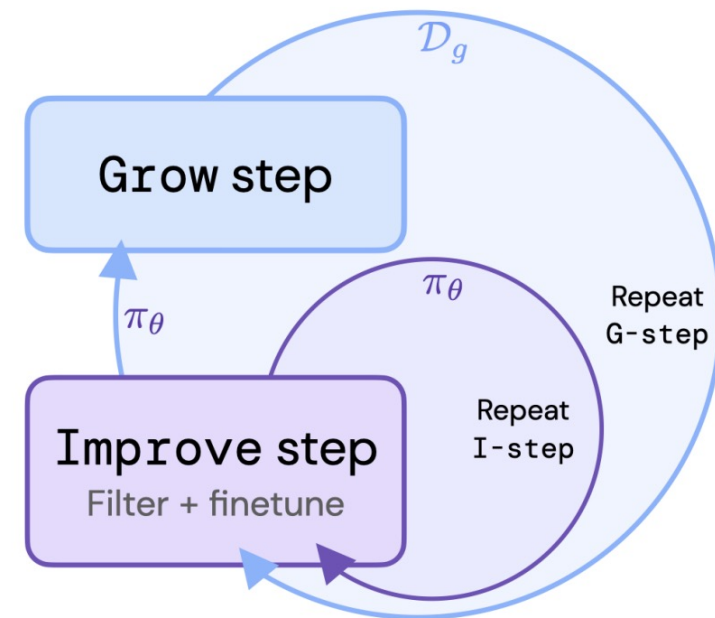


Figure 1 | **ReST method.** During Grow step, a policy generates a dataset. At Improve step, the filtered dataset is used to fine-tune the policy. Both steps are repeated, Improve step is repeated more frequently to amortise the dataset creation cost.

Source: <https://arxiv.org/abs/2308.08998>

# DPO

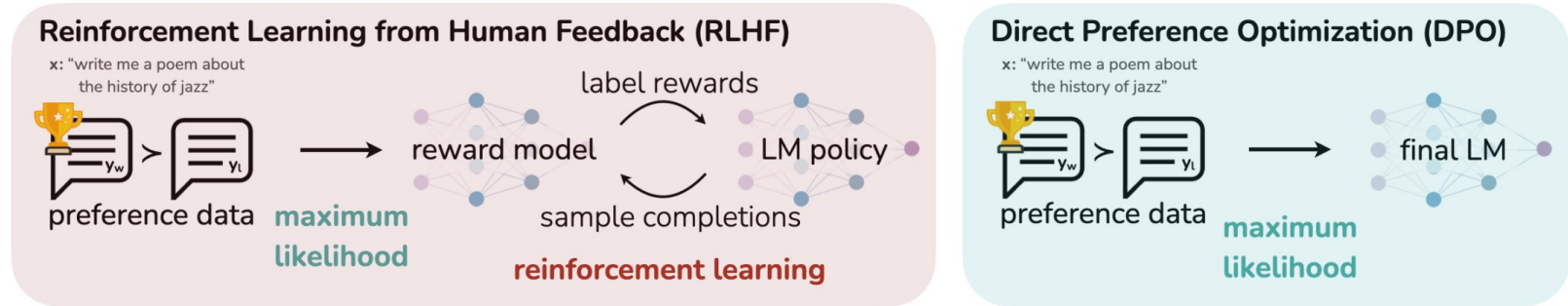


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, without an explicit reward function or RL.

Source: <https://arxiv.org/abs/2305.18290>

## Direct Preference Optimisation (DPO)

- Discard the need for a reward model.
- Computationally cheap: no need for RL (in training prompt gen.)

## Steps

0. Acquire data: multiple response same prompt with human eval.
1. Learn preference model via supervised classification

# DPO

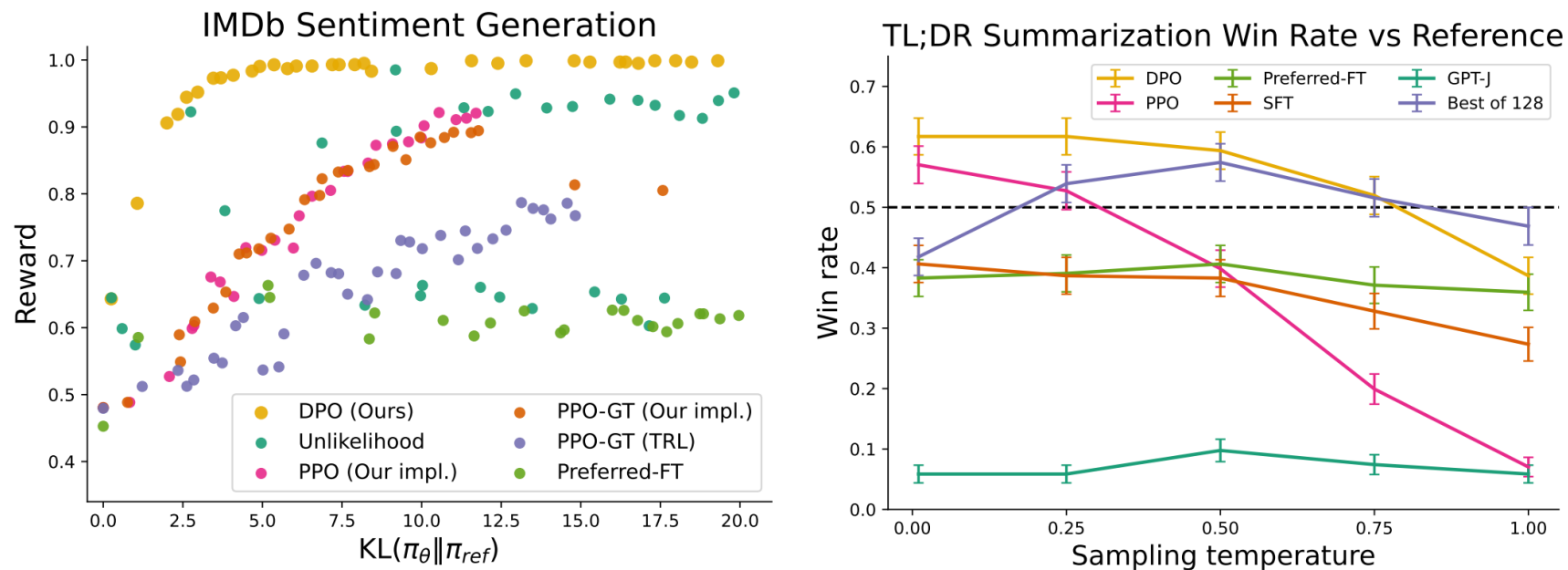


Figure 2: **Left.** The frontier of expected reward vs KL to the reference policy. DPO provides the highest expected reward for all KL values, demonstrating the quality of the optimization. **Right.** TL;DR summarization win rates vs. human-written summaries, using GPT-4 as evaluator. DPO exceeds PPO's best-case performance on summarization, while being more robust to changes in the sampling temperature.

Source: <https://arxiv.org/abs/2305.18290>

## Experimental Evaluation

- DPO on par with RLHF (PPO)

# LLaVA-RLHF

## Aligning Large Multimodal Models with Factually Augmented RLHF

Zhiqing Sun<sup>\*</sup>, Sheng Shen<sup>\*</sup>, Shengcao Cao<sup>\*</sup>,

Haotian Liu, Chunyuan Li, Yikang Shen,

Chuang Gan<sup>†</sup>, Liang-Yan Gui<sup>†</sup>, Yu-Xiong Wang<sup>†</sup>, Yiming Yang<sup>†</sup>, Kurt Keutzer<sup>†</sup>, Trevor Darrell<sup>†</sup>,

► UC Berkeley ► CMU ► UIUC ► UW–Madison ► Microsoft Research ► MIT-IBM Watson AI Lab

<sup>\*</sup>Equal Contribution, <sup>†</sup>Equal Advising

arXiv

Code

Demo

Dataset (RM)

Dataset (SFT)

MMHal-Bench

Model (13b)

Model (7b)

LLaVA-RLHF represents the first open-source **RLHF-trained** large multimodal model for general-purpose visual and language understanding, achieving impressive visual reasoning and perception capabilities mimicking spirits of the multimodal GPT-4 and setting a new state-of-the-art accuracy on LLaVA-Bench, MMBench, and MMHal-Bench.

We propose a new alignment algorithm called **Factually Augmented RLHF (Fact-RLHF)** that augments the reward model with additional factual information such as image captions and ground-truth multi-choice options, which alleviates the reward hacking phenomenon in RLHF and further improves the performance.

LLaVA-RLHF combines a CLIP vision encoder and Vicuna, is fine-tuned with high-quality vision instruction tuning data and Fact-RLHF, and is presented to be **more helpful** and **less hallucinated** than LLaVA or other open-sourced LMMs.

Source: <https://llava-rlhf.github.io/>

**Thank you!**