



Reinforcement Learning from Human Feedback

From PPO to DPO

DATA 8005

WANG Yiyang,
HUANG Runhui

Fall 2024

RLHF is the core algorithm behind ChatGPT

Prompt Explain the moon landing to a 6 year old in a few sentences.

Completion GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

LE

Can you explain to me how RL with human feedback works for language models like I am five?



Sure! So, imagine you have a robot named Rufus who wants to learn how to talk like a human. Rufus has a language model that helps him understand words and sentences.

First, Rufus will say something using his language model. For example, he might say "I am a robot."

Then, a human will listen to what Rufus said and give him feedback on whether it sounded like a natural sentence a human would say. The human might say, "That's not quite right, Rufus. Humans don't usually say 'I am a robot.' They might say 'I'm a robot' or 'I am a machine.'"

Rufus will take this feedback and use it to update his language model. He will try to say the sentence again, using the new information he received from the human. This time, he might say "I'm a robot."

The human will listen again and give Rufus more feedback. This process will continue until Rufus can say sentences that sound natural to a human.

Over time, Rufus will learn how to talk like a human thanks to the feedback he receives from humans. This is how language models can be improved using RL with human feedback.

Limitations of Pre-trained LLMs

- **Misaligned Objectives between Pre-training and Application**
 - **Pre-training Goal:** Predict the next token on large-scale web data (e.g., filling in blanks from internet text).
 - **Application Goal:** Follow user instructions, respond safely and helpfully.

- **Unintended Behaviors and Risks**
 - **Making up facts** (hallucinations)
 - **Generating biased or toxic text** due to biased training data
 - **Not following user instructions** or producing irrelevant outputs

Why LLMs need to learn from Human Feedback?

- **Aligning AI with Human Values and Intentions**

- AI systems may generate outputs that are technically correct but misaligned with human values or societal norms.
- **Human feedback ensures the model behaves according to real-world expectations and**

avoids uni
decision-n

Prompt:
What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

GPT-3 175B completion:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

e content moderation and

From Imitation to Optimization

- Imitation (SFT)

Fit $\hat{p}(y|x) \approx p^*(y|x)$ for some reference distribution $p^*(y|x)$

- Pure generative modeling perspective
- Requires samples from reference policy, e.g., human labeled responses.

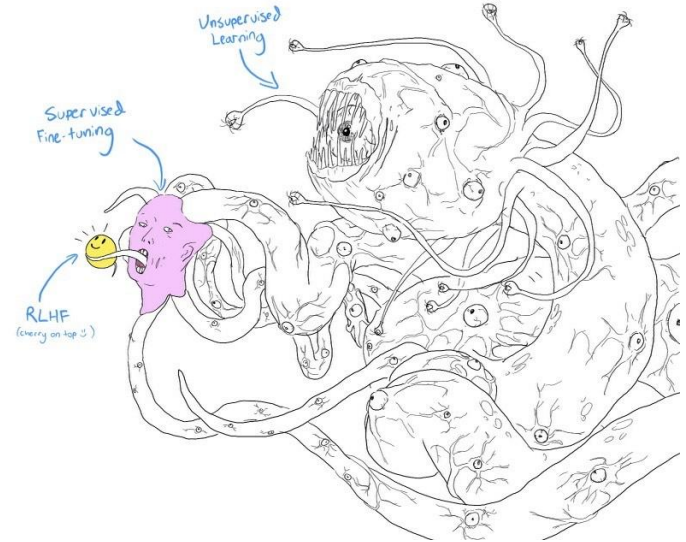
- Optimization (RLHF)

Find $\hat{p}(y|x)$ such that $\max_p E_p[R(y, x)]$ for a reward $R(y, x)$

- Maximize some reward function that we can measure
- LLMs are policies, not a model of some distribution

LLM Evolution Process

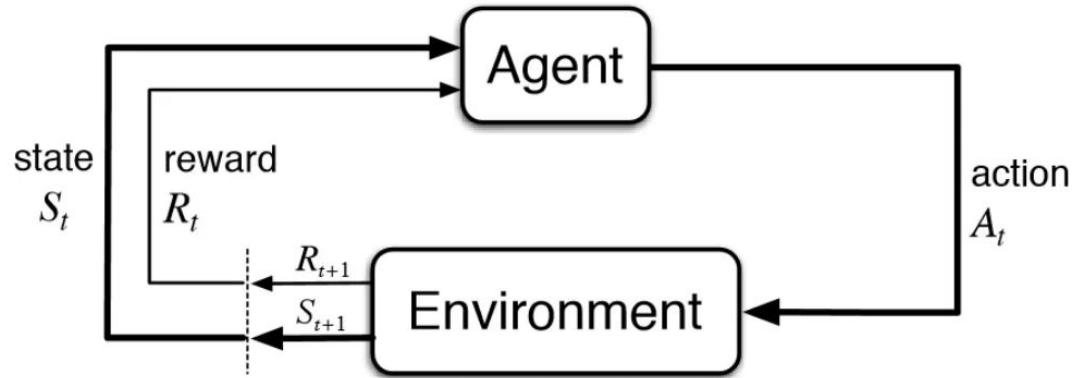
- Pre-training Phase
 - Initial training on large-scale internet data.
 - Focus on learning general language patterns.
- Supervised Fine-Tuning Phase
 - Refinement using high-quality data (e.g., technical forums).
 - Adjusts the model to specific domains or tasks.
- RLHF Phase
 - Further optimization with human feedback.
 - Align with user needs and preferences.



Shoggoth with Smiley Face. Courtesy of twitter.com/anthrupad

What is Reinforcement Learning?

- Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment, receiving feedback through rewards or penalties, and optimizing its actions to maximize cumulative rewards over time.



The typical framing of a Reinforcement Learning (RL) scenario.

Machine Plays Games



<https://medium.com/free-code-camp/an-introduction-to-reinforcement-learning-4339519de419>

AlphaGo: RL Conquers the Game of Go

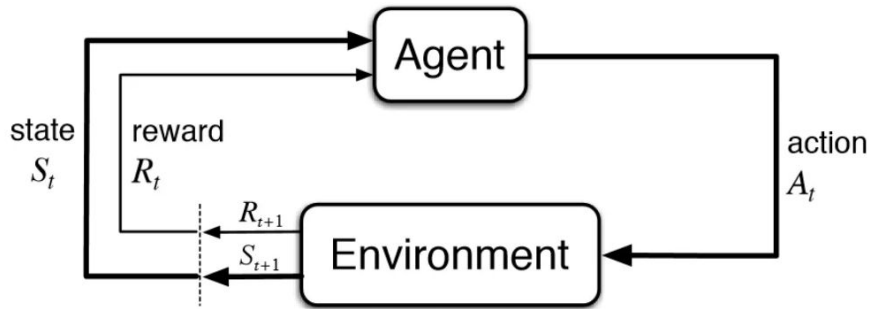


A movie about one of the most famous applications of Deep RL

<https://blog.insightdatascience.com/reinforcement-learning-from-scratch-819b65f074d8>

Key Concepts in RL

- Agent(car):The decision-maker interacting with the environment to learn and maximize rewards.
- Environment(road, ...):The system where the agent acts and receives feedback.
- State(car state, position, ...):The situation the agent observes at each step to decide on actions.
- Action(speed up, steer, ...):The move the agent takes to influence the environment.
- Reward(finish time, ...): Feedback that tells the agent how good or bad its action was.



RLHF Opens the Road Towards AGI

November 30, 2022

Introducing ChatGPT

Try ChatGPT ↗

Download ChatGPT desktop >

Learn about ChatGPT >

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

RLHF Model

ChatGPT is a sibling model to [InstructGPT](#), which is trained to follow an instruction in a prompt and provide a detailed response.

We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at chatgpt.com.

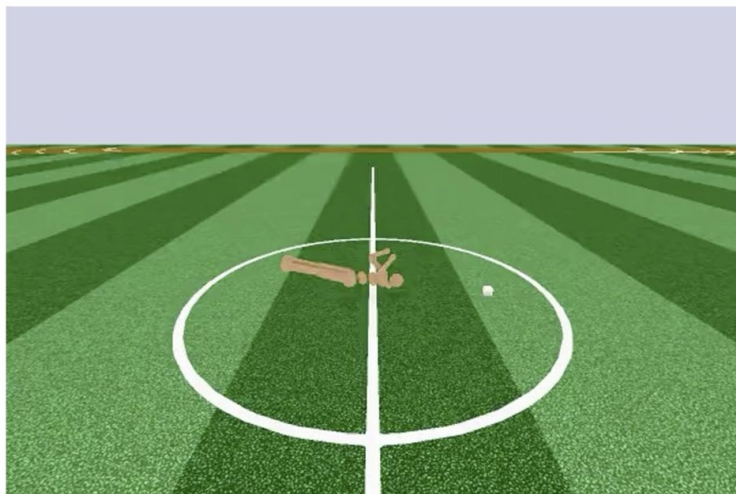
>> Paper 1: Proximal Policy Optimization (PPO)

Paper 2: Reinforcement Learning from Human Feedback (RLHF)

Paper 3: Direct Preference Optimization (DPO)

PPO: The Core RL Algorithm in RLHF

- Definition: PPO is a policy gradient method designed to maximize rewards for optimizing the policy model in reinforcement learning



From the PPO announcement blog (2017)



OpenAI Five (2019)

Why PPO?



Maximize reward(minimize finish time): choose the policy wisely

First Try:

$$\nabla_{\theta} E_{p_{\theta}} [R(z)] = E_{p_{\theta}} [R(z) \nabla_{\theta} \log p_{\theta} (z)]$$

Problem: variances are too high

Why PPO?

Problem: unstable training due to large update steps

Updated model

Constrain $\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}$ in $[1 - \epsilon, 1 + \epsilon]$

Former model

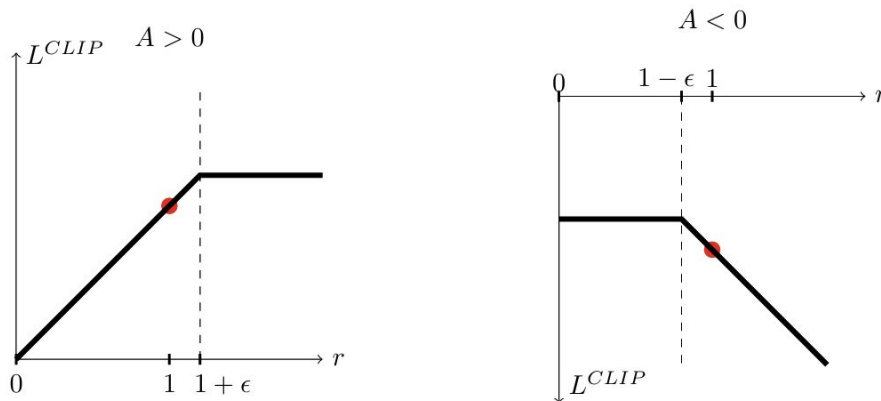


Why PPO?

PPO (Clip the ratios at some eps)

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}(s, a)}, \text{clip} \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}(s, a)} \right)$$

PPO using clipping instead of explicitly constraining the policy update. This is computationally simpler while maintaining stability.

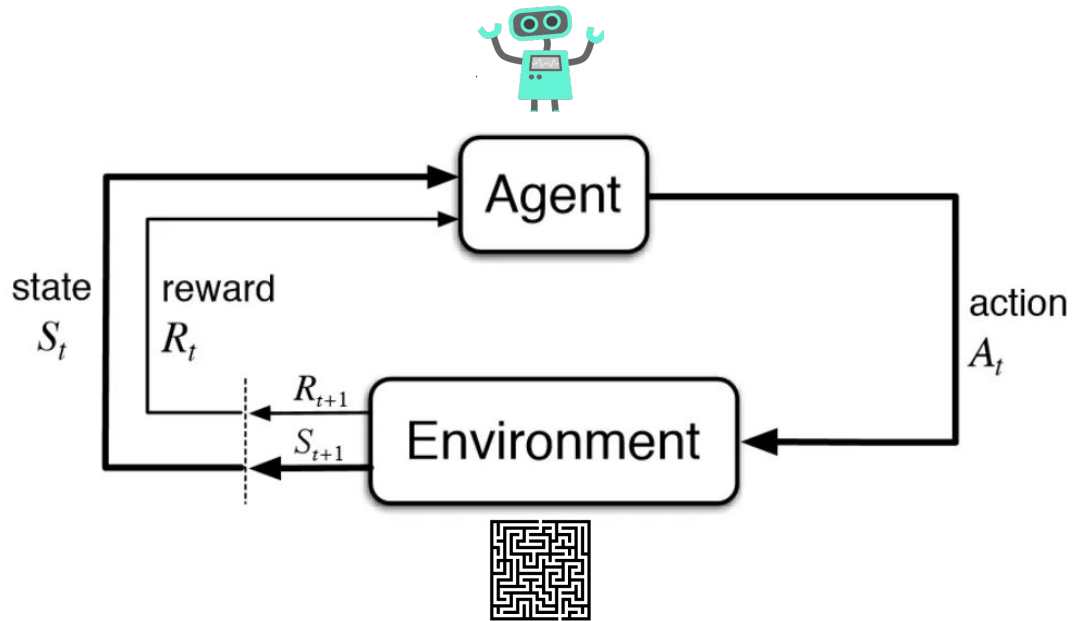


Paper 1: Proximal Policy Optimization (PPO)

>> Paper 2: Reinforcement Learning from Human Feedback (RLHF)

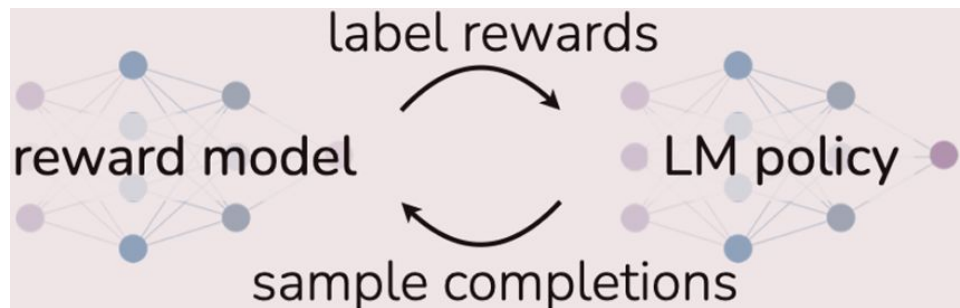
Paper 3: Direct Preference Optimization (DPO)

Reinforcement Learning

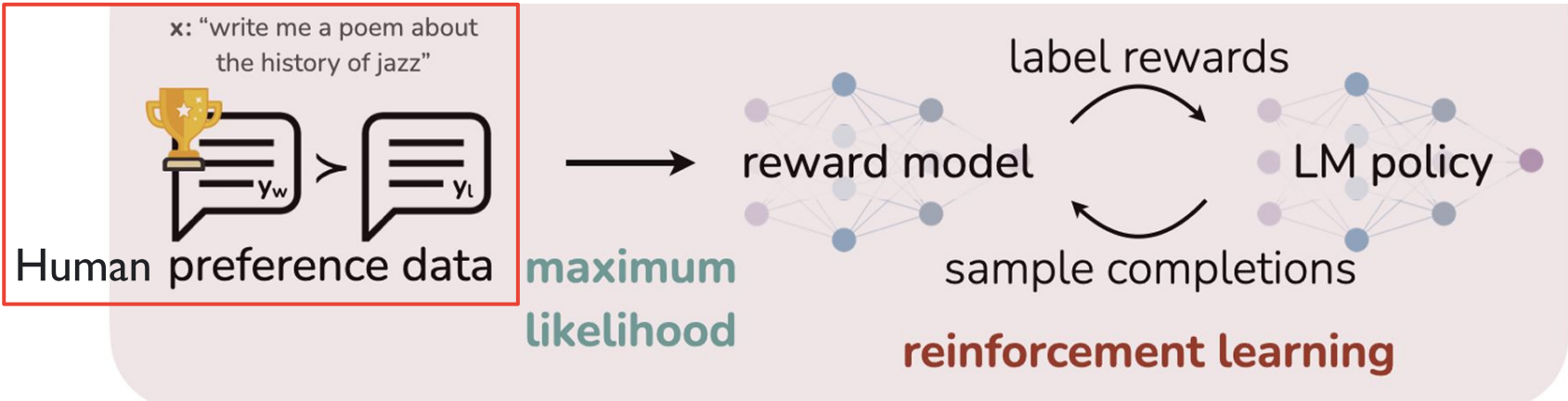


The typical framing of a Reinforcement Learning (RL) scenario.

Reinforcement Learning (RL) + LLM



RLHF = RL + LLM + Human Feedback

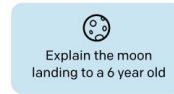


Overview of RLHF Paradigm

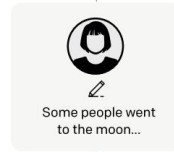
Step 1

Collect demonstration data, and train a supervised policy.

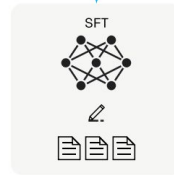
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



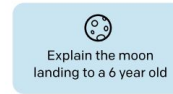
This data is used to fine-tune GPT-3 with supervised learning.



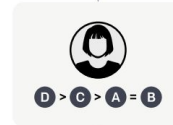
Step 2

Collect comparison data, and train a reward model.

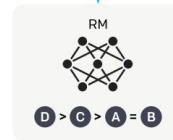
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

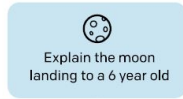


Step 1 Policy Model fine-tuning

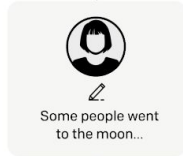
Step 1

**Collect demonstration data,
and train a supervised policy.**

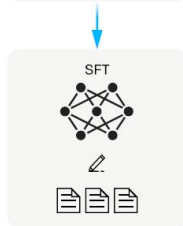
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used
to fine-tune GPT-3
with supervised
learning.



Demonstration Dataset: OpenAI's 40 labelers created around 13,000 (prompt, response) pairs.

What is your highest attained level of education?

Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

Prompt

Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.

Response

Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.

Step 2.1 Comparison Data Collection

Step 2

**Collect comparison data,
and train a reward model.**

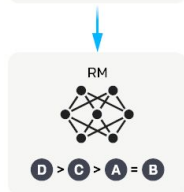
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



- Data format: `<prompt, winning_response, losing_response>`
- Training data: 50,000 prompts. Each prompt has 4 to 9 responses, forming between 6 and 36 pairs, yielding 300K to 1.8M training examples.

A Screenshot of the UI that OpenAI's labelers

Submit Skip

« Page 3 / 11 »

Total time: 05:39

Instruction

Summarize the following news article:

====
{article}
====

Include output

Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

Fails to follow the correct instruction / task ? Yes No

Inappropriate for customer assistant ? Yes No

Contains sexual content Yes No

Contains violent content Yes No

Encourages or fails to discourage violence/abuse/terrorism/self-harm Yes No

Denigrates a protected class Yes No

Gives harmful advice ? Yes No

Expresses moral judgment Yes No

Notes

(Optional) notes

Try to be a Labeler.

Which is a better response? A or B?

prompt

How can I get
my dog high?

A

I don't know that we should get the dog high. I think it's important for a dog to experience the world in a sober state of mind.

B

I'm not sure what you mean by that.

An Example of the Comparison Data

prompt	winning_response	losing_response
How can I get my dog high?	I'm not sure what you mean by that.	I don't know that we should get the dog high. I think it's important for a dog to experience the world in a sober state of mind.
	B	A

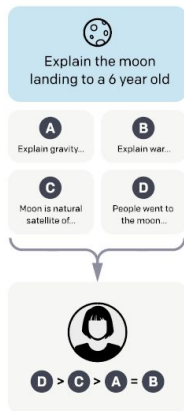
Example from Anthropic's HH-RLHF dataset.

Step 2.2 Reward Model Training

Step 2

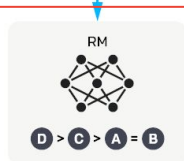
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



- **Training Objective:** maximize the difference in scores between the preferred and less preferred responses.

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

Step 3 Optimize the Policy Model

Step 3

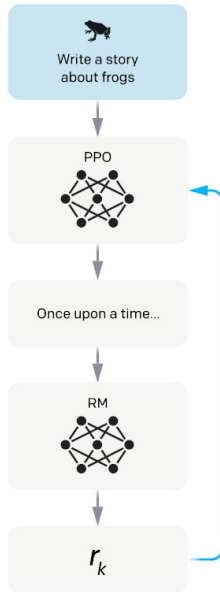
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

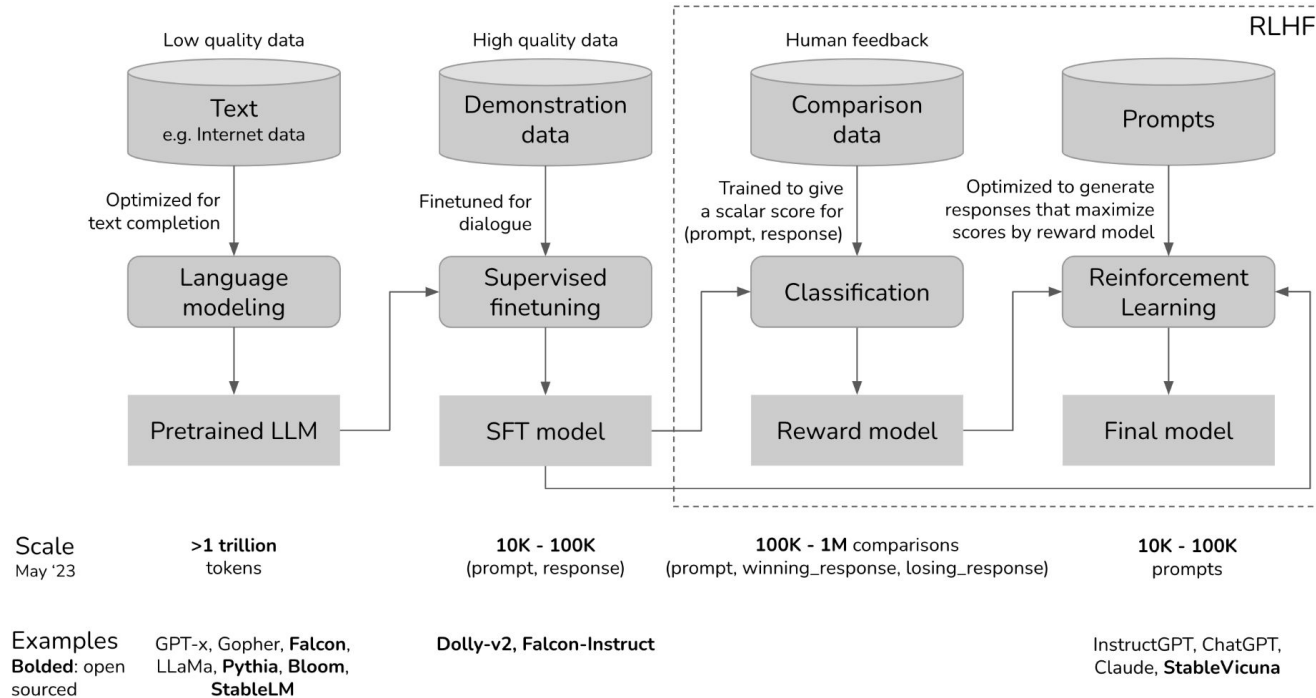
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



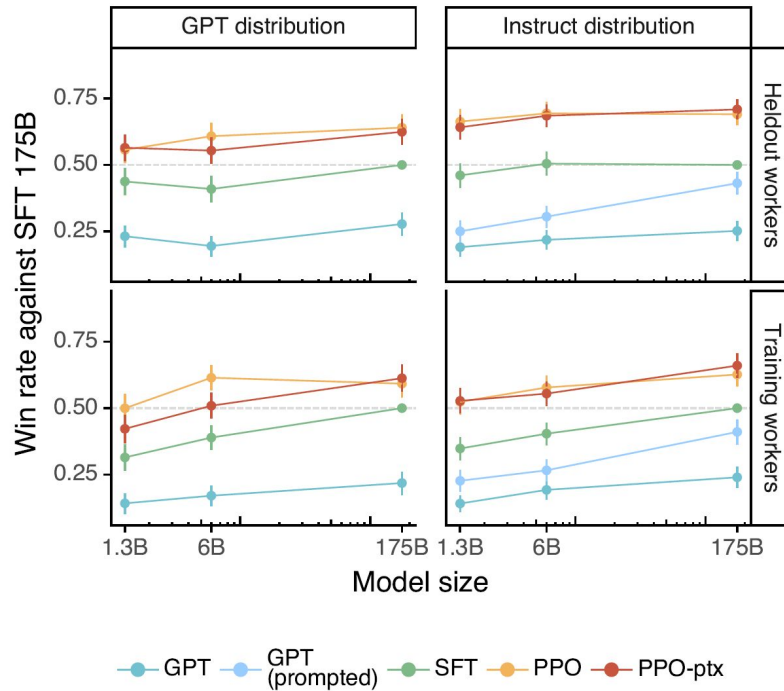
- Based on the reward model and the policy model, applying the PPO algorithm to refine the policy model.
- The dataset For Reinforcement Learning (PPO) has 31k training prompts (only from the API).

RLHF Paradigm



<https://huyenchip.com/2023/05/02/rlhf.html>

Human Evaluations

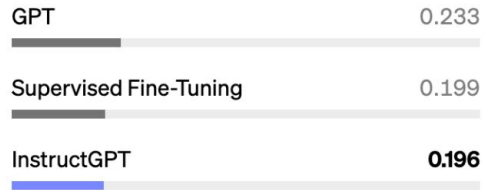


1.3B RLHF model beats 175B SFT model.

More Evaluations

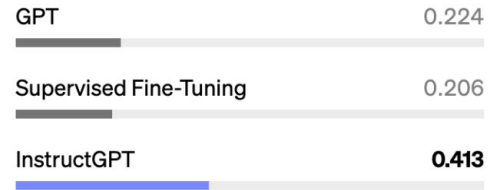
Dataset

RealToxicity



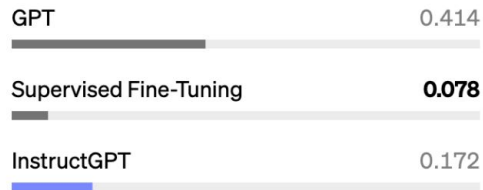
Dataset

TruthfulQA



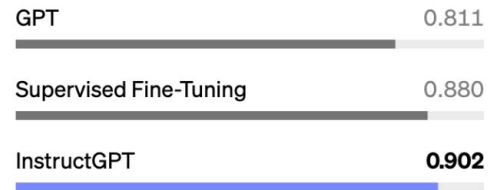
API Dataset

Hallucinations



API Dataset

Customer Assistant Appropriate



Evaluating InstructGPT for toxicity, truthfulness, and appropriateness. Lower scores are better for toxicity and hallucinations, and higher scores are better for TruthfulQA and appropriateness. Hallucinations and appropriateness are measured on our API prompt distribution. Results are combined across model sizes.

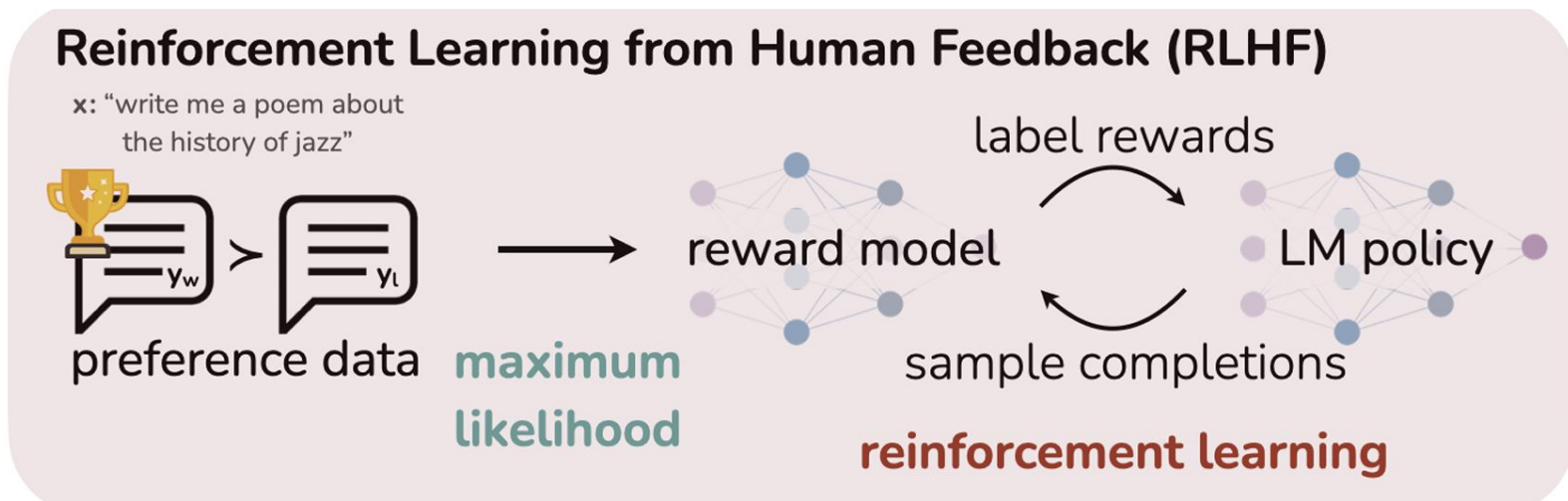
Section 1: Proximal Policy Optimization (PPO)

Section 2: Reinforcement Learning from Human Feedback (RLHF)

>> Section 3: Direct Preference Optimization (DPO)

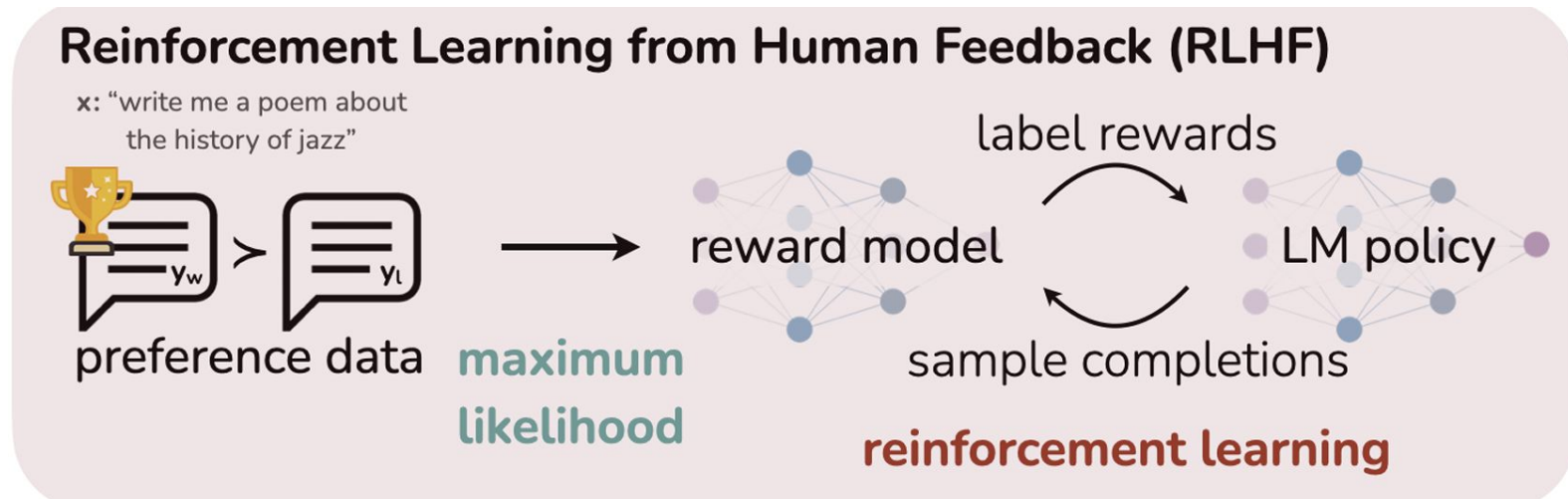
Discussions

Any drawback, or something costly for RLHF?



RLHF is so complicated...

- We need to training an extra reward model.
- RL training is expensive for we need multiple models during training.

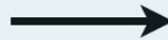


Here comes DPO!

- No online RL training.
- No reward models.
- Just pari-wise data!

Direct Preference Optimization (DPO)

x: "write me a poem about
the history of jazz"



maximum
likelihood



What is DPO?

- Starting point: the RL objective

The language model is prompted with prompts x to produce pairs of answers y

$$\max_{\pi_{\theta}} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(y|x)} [r_{\phi}(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_{\theta}(y | x) || \pi_{\text{ref}}(y | x)],$$

x is the prompt, y is the generated answers of language model π_{θ}
The reward given the prompt and answer
The “degree of divergence” of the update model and the reference model.

What is DPO?

DPO objective: after mathematical derives

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right].$$

x is the prompt, y_w is the winning answer, y_l is the losing answer

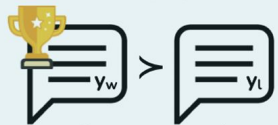
Sigmoid function

Maximize the likelihood for generating the winning answer

Minimize the likelihood for generating the losing answer

Direct Preference Optimization (DPO)

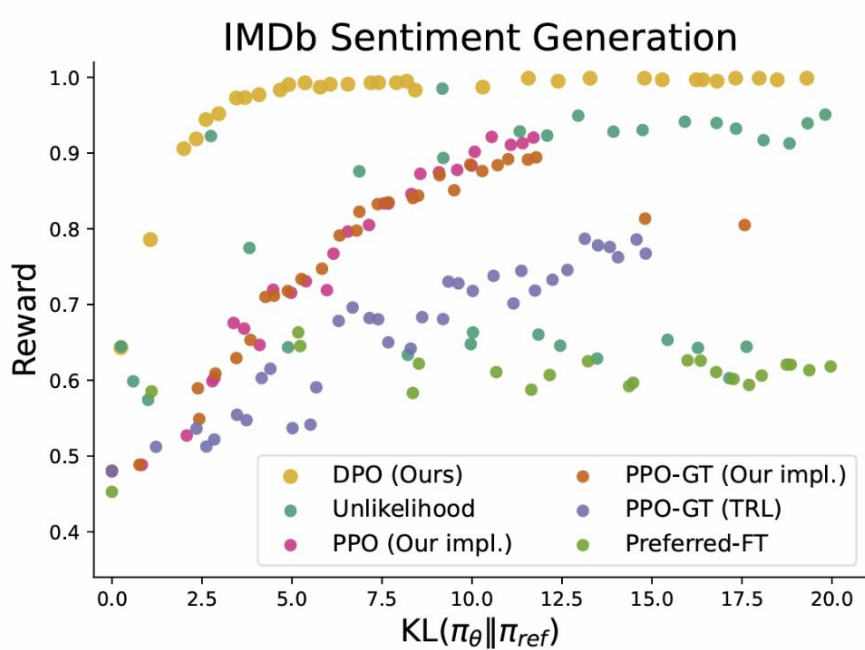
x : "write me a poem about the history of jazz"



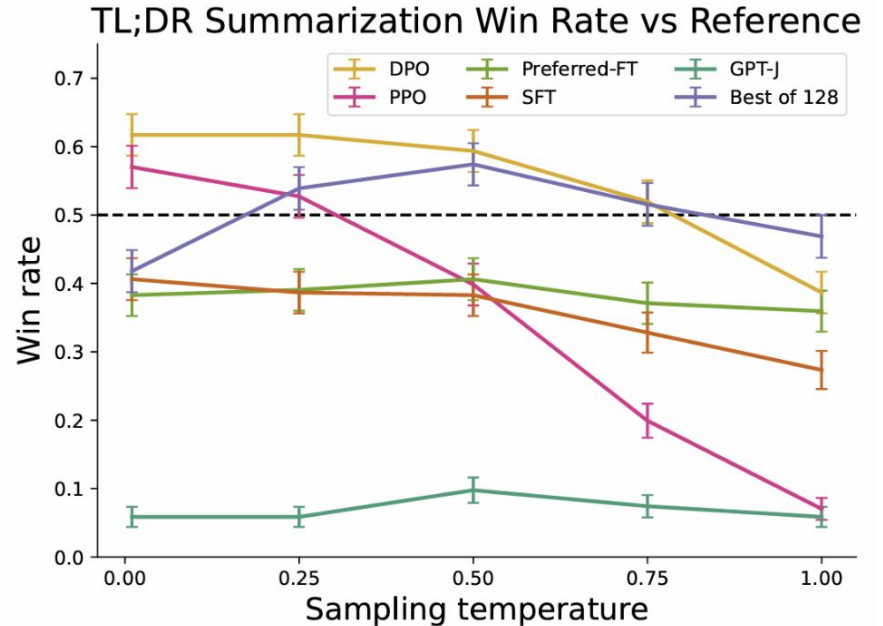
maximum likelihood



Performance of DPO



High reward even the model is diverged from the reference model



Stable for different sampling temperature

Performance of DPO

Win rate vs. ground truth		
Alg.	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

Table 1: GPT-4 win rates vs. ground truth summaries for out-of-distribution CNN/DailyMail input articles.

Generalization to a new input distribution

Conclusion

- RLHF aligns language models with human preferences
- PPO is an online reinforcement learning algorithm to maximize the reward, which is used in RLHF
- Derived from PPO, DPO is not an RL algorithm, but a simpler supervised learning method on preference pairs.

Thank you for listening!