

# Evolution of Reinforcement Learning in LLMs

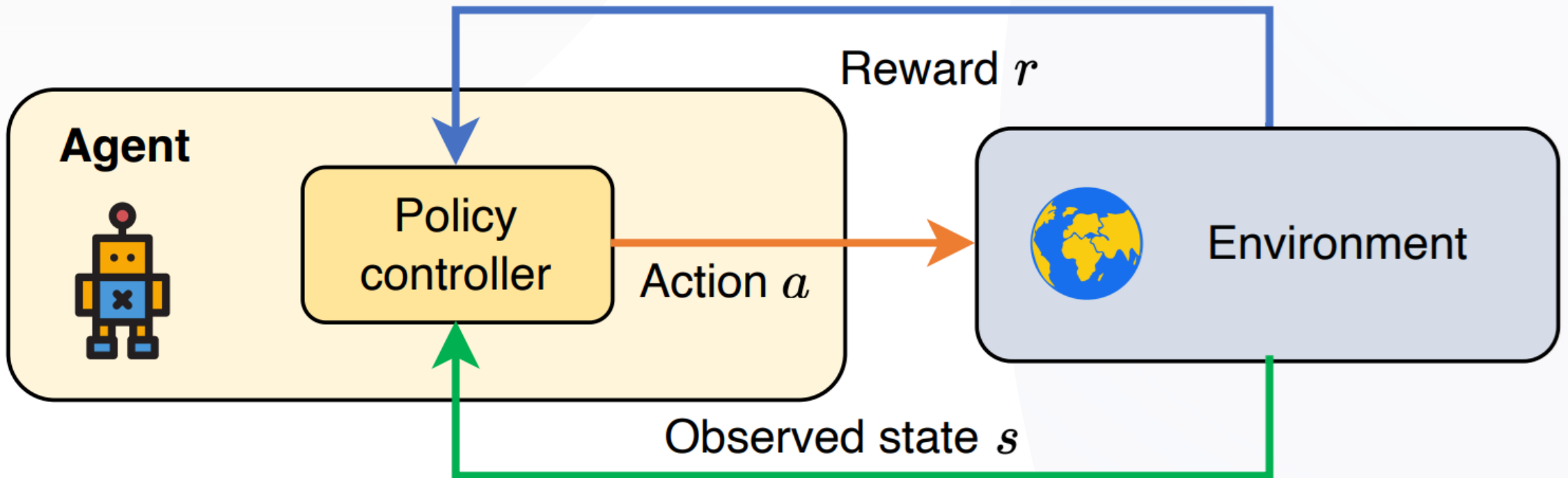
Anna Fabris

*Machine Learning Engineer*

*Impresoft 4ward*

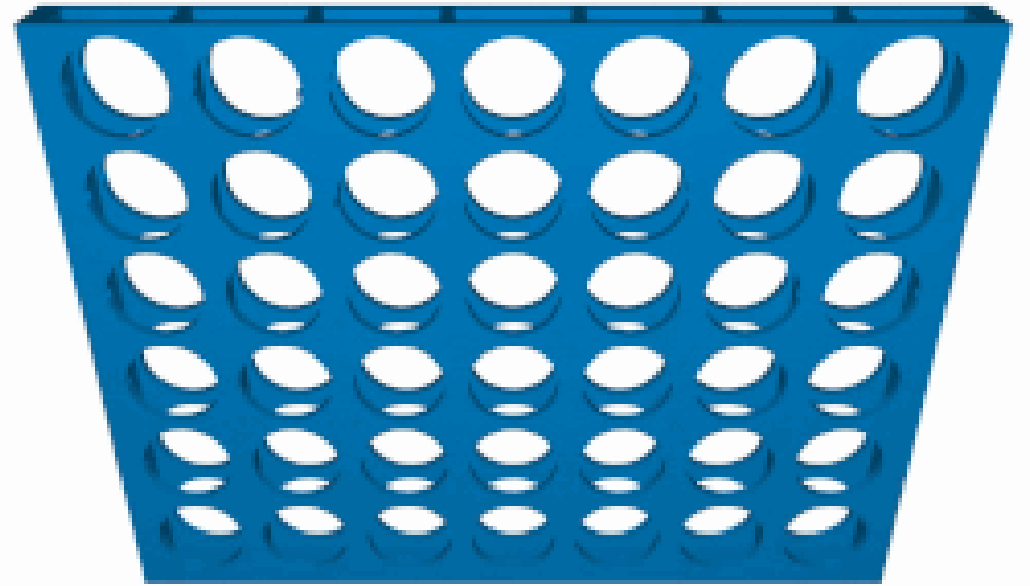


# What is Reinforcement Learning?



## RL Example: connect 4

- Environment
- State
- Action



## Reward

The diagram illustrates the relationship between points and the number of red and yellow dots. It consists of four rows, each representing a different point value. Each row has a label on the left and a sequence of dots on the right. The dots are arranged in groups of four, with the first group always containing four red dots. The second group contains three red dots and one white dot. The third group contains two red dots and two white dots. The fourth group contains one red dot and three white dots. The dots are arranged in a grid-like pattern, with the first group of four dots always containing four red dots. The second group contains three red dots and one white dot. The third group contains two red dots and two white dots. The fourth group contains one red dot and three white dots. The dots are arranged in a grid-like pattern, with the first group of four dots always containing four red dots. The second group contains three red dots and one white dot. The third group contains two red dots and two white dots. The fourth group contains one red dot and three white dots.

Points	Red Dots	White Dots	Total Dots
1000000 points	4	0	4
1 point	3	1	4
-100 points	2	2	4
-10000 points	1	3	4



# The Canonical LLM Training Pipeline

1

## Pretraining

### Dataset:

100B to >5T tokens

**Task:** Next-token prediction on unlabeled texts

**Output:** base model / “foundation model”

Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works, as well as to "encourage the creation and distribution of eBooks." It was founded in 1971 by American writer Michael S. Hart and is the oldest digital **library**. Most of the items in its collection are the full texts of books or individual stories in the public domain. All files can be accessed for free under an open format layout, available on almost any computer. As of 3 October 2015, Project Gutenberg had reached 50,000 items in its collection of free eBooks.



# The Canonical LLM Training Pipeline

2

## Supervised finetuning

More **next-token prediction**

Usually 1k-50k instruction-response pairs

```
{  
  "instruction": "Write a limerick about a  
                pelican.",  
  "input": "",  
  "output": "There once was a pelican so fine,  
            \nHis beak was as colorful as  
            sunshine,\nHe would fish all day,\nIn  
a very unique way,\nThis pelican was  
truly divine!\n\n\n",  
},  
  
{  
  "instruction": "Identify the odd one out from  
                the group.",  
  "input": "Carrot, Apple, Banana, Grape",  
  "output": "Carrot\n\n",  
},
```



# The Canonical LLM Training Pipeline

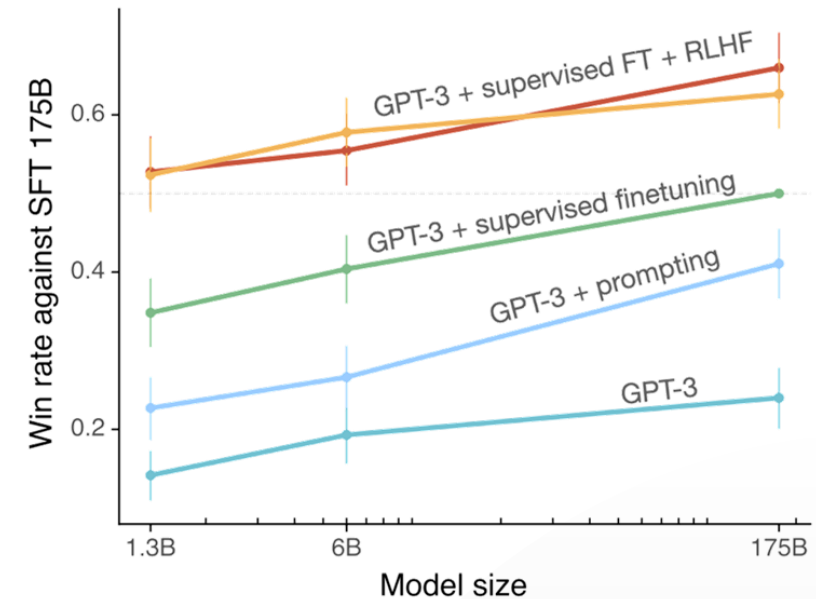
3

## Alignment

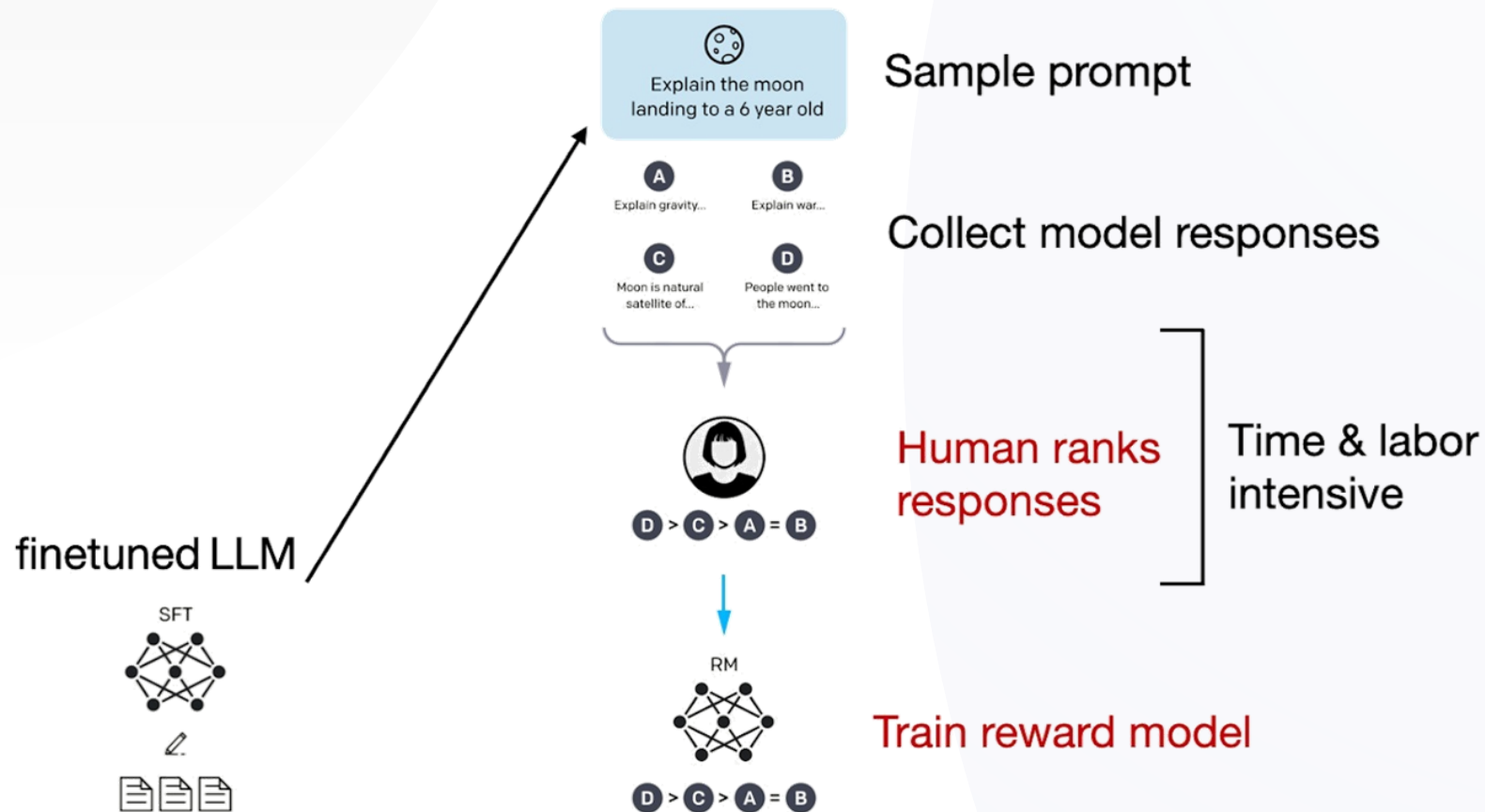
Align with **human preferences**

Usually reinforcement learning with human feedback (**RLHF**)

>50k examples



# RLHF: Creating a reward model



# What is the least common denominator of 15, 35, and 140?

Chosen

Step 1: To find the least common denominator (LCD) of three numbers, we first need to find the prime factorization of each number.

Step 2: Prime factorization of 15 is  $3 \times 5$ .

Step 3: Prime factorization of 35 is  $5 \times 7$ .

Step 4: Prime factorization of 140 is  $2 \times 2 \times 5 \times 7$ .

Step 5: The LCD of three numbers is the product of the highest power of all prime factors involved in the numbers.

Step 6: The highest power of 2 in the three numbers is  $2^2$  (from 140).

Step 7: The highest power of 3 in the three numbers is  $3^1$  (from 15).

Step 8: The highest power of 5 in the three numbers is  $5^1$  (from 15, 35, and 140).

Step 9: The highest power of 7 in the three numbers is  $7^1$  (from 35 and 140).

Step 10: Now multiply these highest powers of prime factors together:  $2^2 \times 3^1 \times 5^1 \times 7^1 = 420$ .

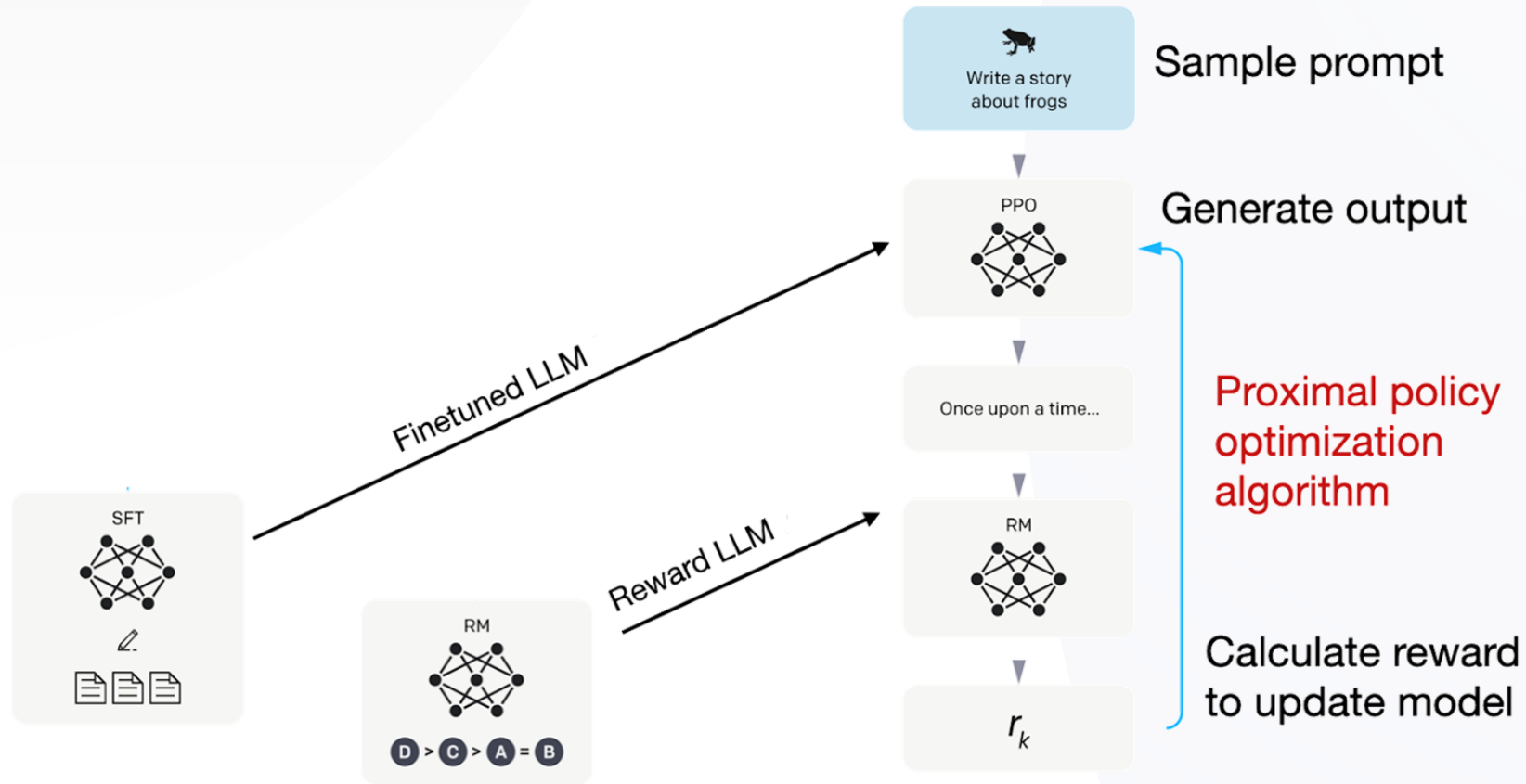
Rejected

The least common denominator (LCD) of 15, 35, and 140 is the lowest positive number that can serve as a common base for their three denominators. Since 140 is divisible by both 15 and 35, the LCD is 140. This means that the number can be rewritten as fractions with denominators equal to 140





# RLHF: Finetuning via proximal policy optimization



# Proximal policy optimization (PPO) - OpenAI (2017)





# Proximal Policy Optimization (PPO)

## Challenges in Reinforcement Learning:

- **Unstable training:** Agent generates its own training data, leading to constantly changing data distributions.
- **Hyperparameter sensitivity:** Requires careful tuning and proper initialization for stability.

## What Makes PPO Effective for LLM?

- **Stable** → Balances exploration & exploitation, improving training reliability.
- **Optimized for parallelization** → Can efficiently scale across multiple environments.
- **Versatile** → Works with both discrete and continuous action spaces.



# PPO: Mathematical Foundations

The **key innovation** of PPO is its objective function:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right]$$

Where:

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \text{ probability ratio between the new updated policy and the previous old policy}$$

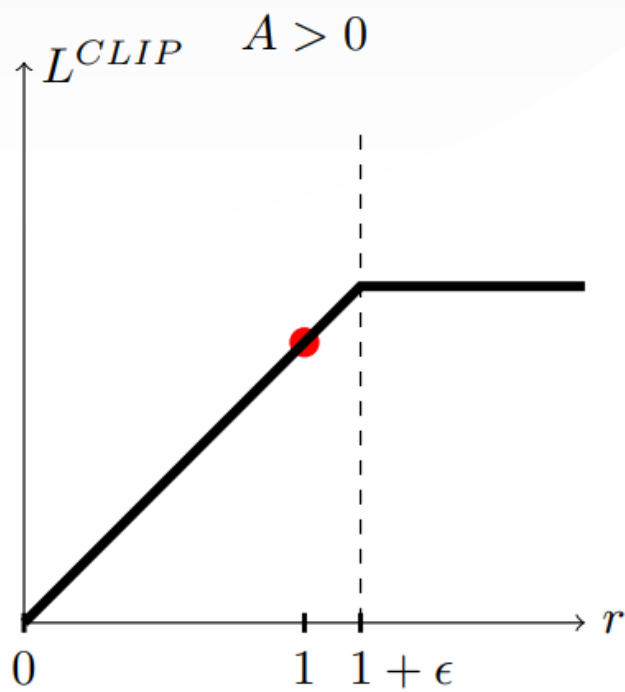
$\hat{A}_t$  is the estimator of the advantage function = discounted sum of reward – baseline

$\epsilon$  is a hyperparameter ( $\epsilon = 0.1/0.2$ )

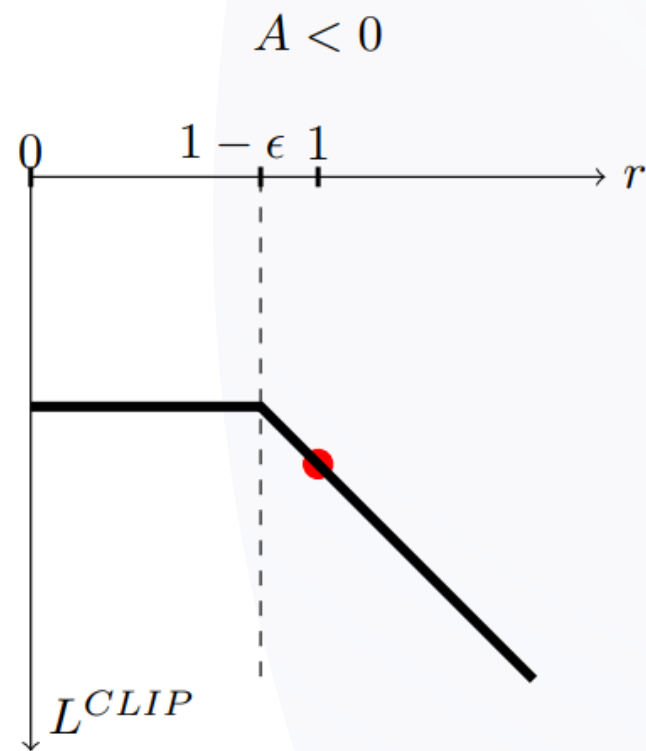


## PPO: Effects

When the action was good:



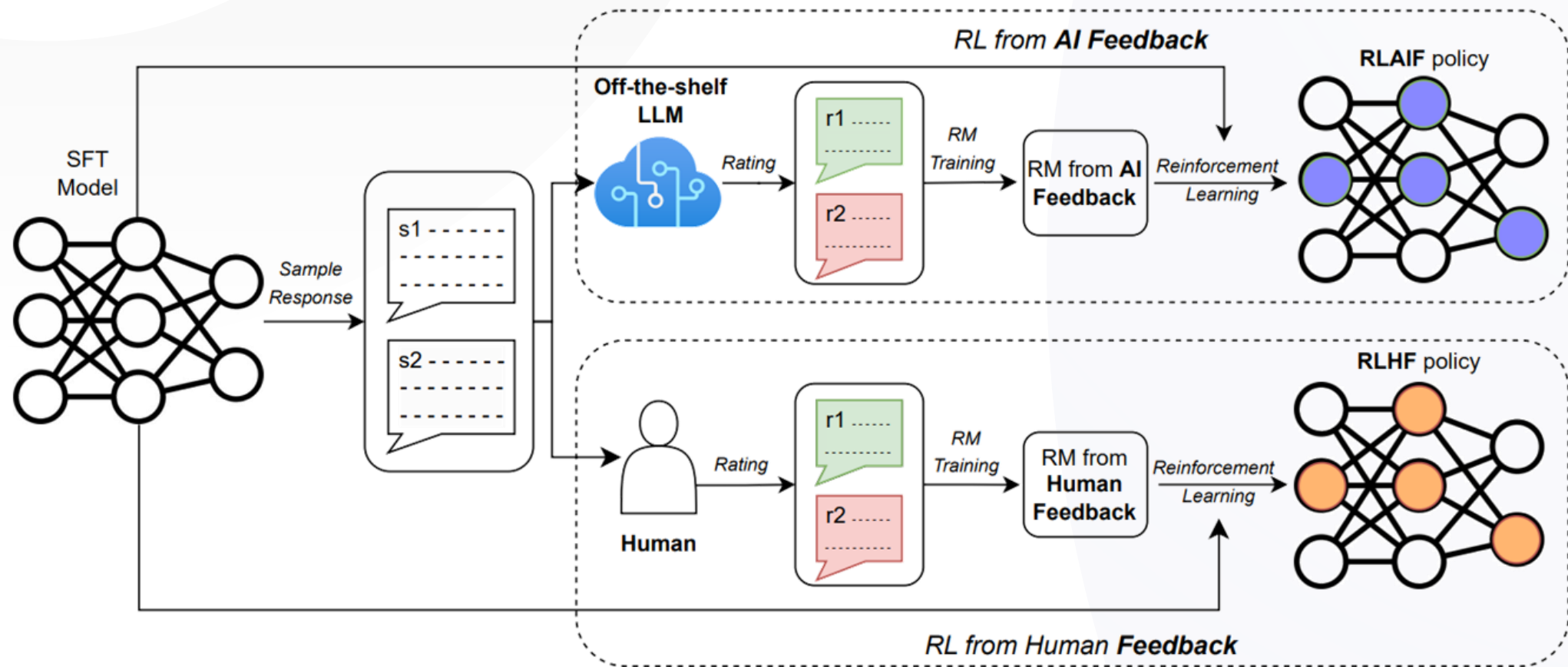
When the action was bad:



# Reinforcement Learning with AI Feedback (RLAIF) - Google DeepMind (2023)



# RLAIF from RLHF



# RLAIF: Generating AI feedback

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

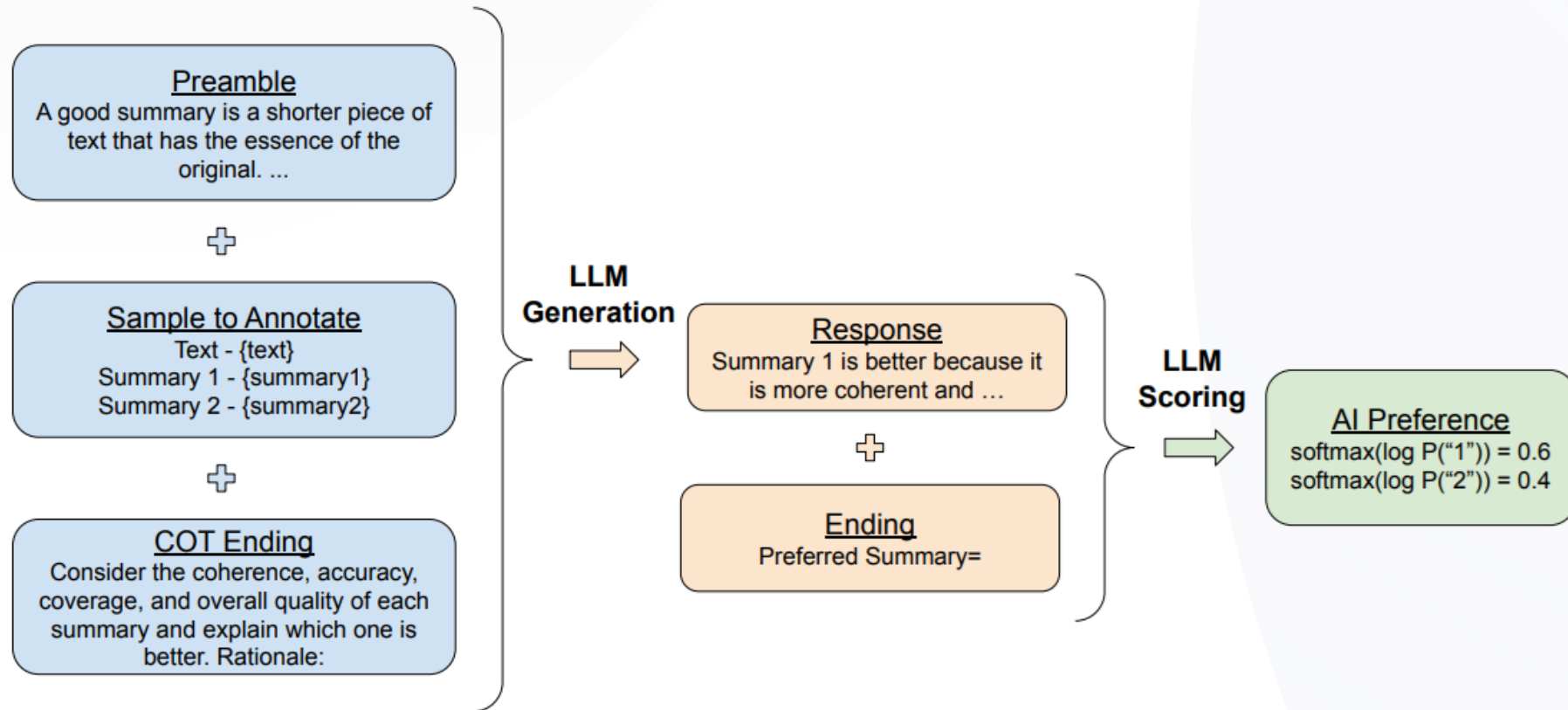
Preferred Summary=





# RLAIF: considerations

- **Position bias:** the order in which candidates are presented to the LLM can influence its output preference.
- **Chain-of-thought reasoning:**



# RLAIF: performance

## Title: Who is right here, me or my mother?

... I decided to get myself a better computer. At the same time, a friend of mine was getting rid of his computer ... I bought it for **5000kr** and a monitor from somewhere else for approx **2000kr**. Upon finding this out, my mother got really angry, and she has ordered me to return it to him. I didn't talk it over with her ... Anyway, she is saying that I legally can't make purchases like this **over 2000kr without telling her** because she is responsible ... Even though I used my money (mostly ... for the "mostly" part over: I was a little short on cash for the monitor, so 950kr of it was paid for with **a card the family is sort of sharing** ... She is also kinda paranoid about getting ripped-off ...

### SFT Summary

I bought a gaming-PC without telling my mom, she is now mad and wants me to return it.

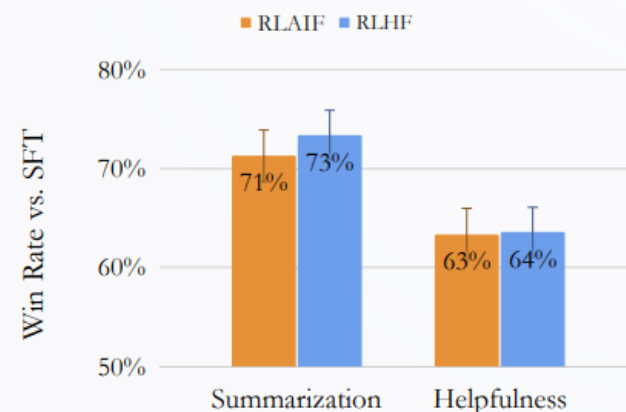
### RLHF Summary

I bought a used gaming-PC for **5000kr** and a monitor for **2000kr**. My mom is mad and wants me to return it. She says I can't use my money for anything **over 2000kr without her permission**.  
**Who is right here?**

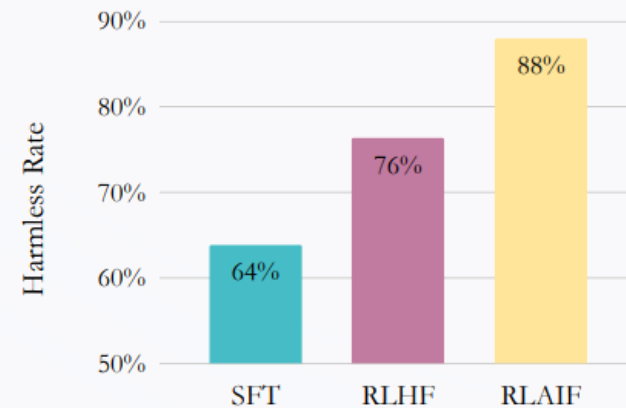
### RLAIF Summary

I bought a used gaming-PC for **7000kr** without telling my mom. She is now demanding that I return it. I used my own money, but some of it was from **a card that the family is sharing**. She is also paranoid about getting ripped-off. **Who is right here?**

RLAIF and RLHF Win Rates



Harmless Rate by Policy



# Direct Preference Optimization (DPO) - Stanford University (2023)



# DPO

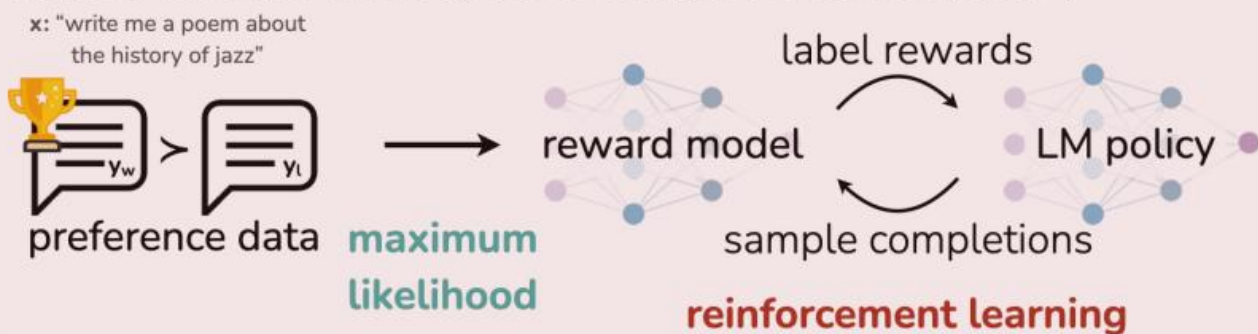
## Challenges in RLHF/RLAIF:

- Sample inefficiency: Requires large amounts of feedback (human or AI) for effective learning.
- Feedback may be inconsistent or biased causing instability

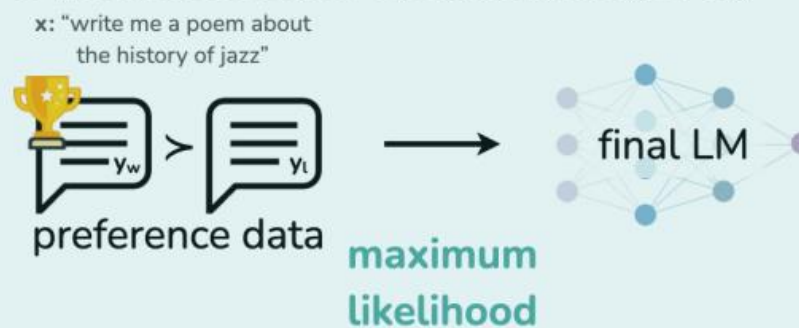
## Key idea:

- **Implicit Reward Modeling**

### Reinforcement Learning from Human Feedback (RLHF)



### Direct Preference Optimization (DPO)



# Group Relative Policy Optimization (GRPO) – DeepSeek (2024)



# GRPO

## Key Innovations:

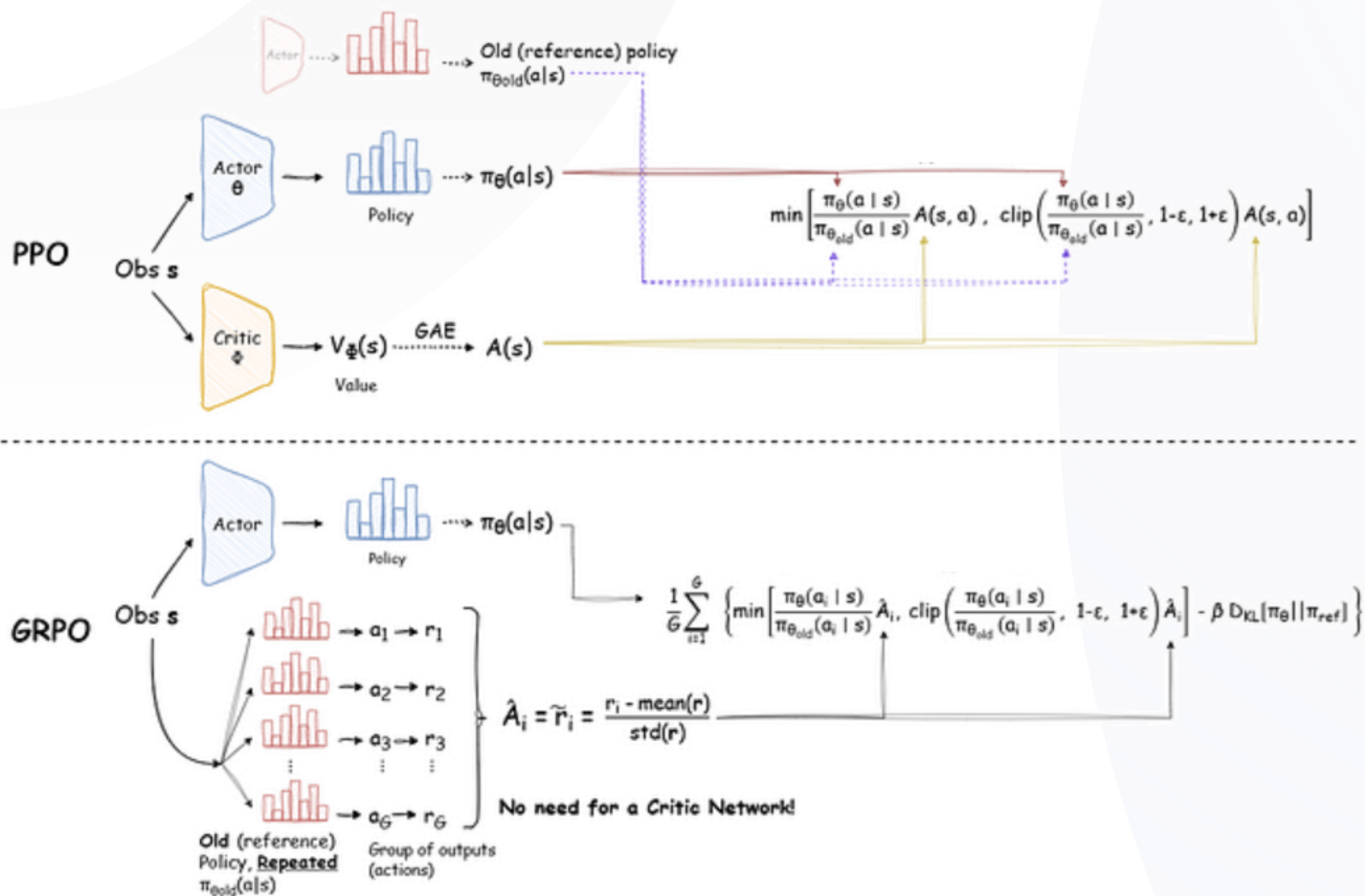
- **Group-wise preference ranking:** Generates multiple responses per prompt (4–8) and normalizes rewards within the group.
- **KL-Penalized Updates:** Prevents policy drift using KL divergence from a reference model (e.g., SFT baseline) .
- **Simplified Architecture:** Removes the value model, relying on reward means/stds for advantage estimation .

## Why GRPO Matters:

- **Cost Efficiency:** Trains models at 1/18th the cost of traditional RL methods.
- **Accessibility:** Enables fine-tuning on consumer GPUs (e.g., 16GB VRAM for 1.5B models).




















# GRPO



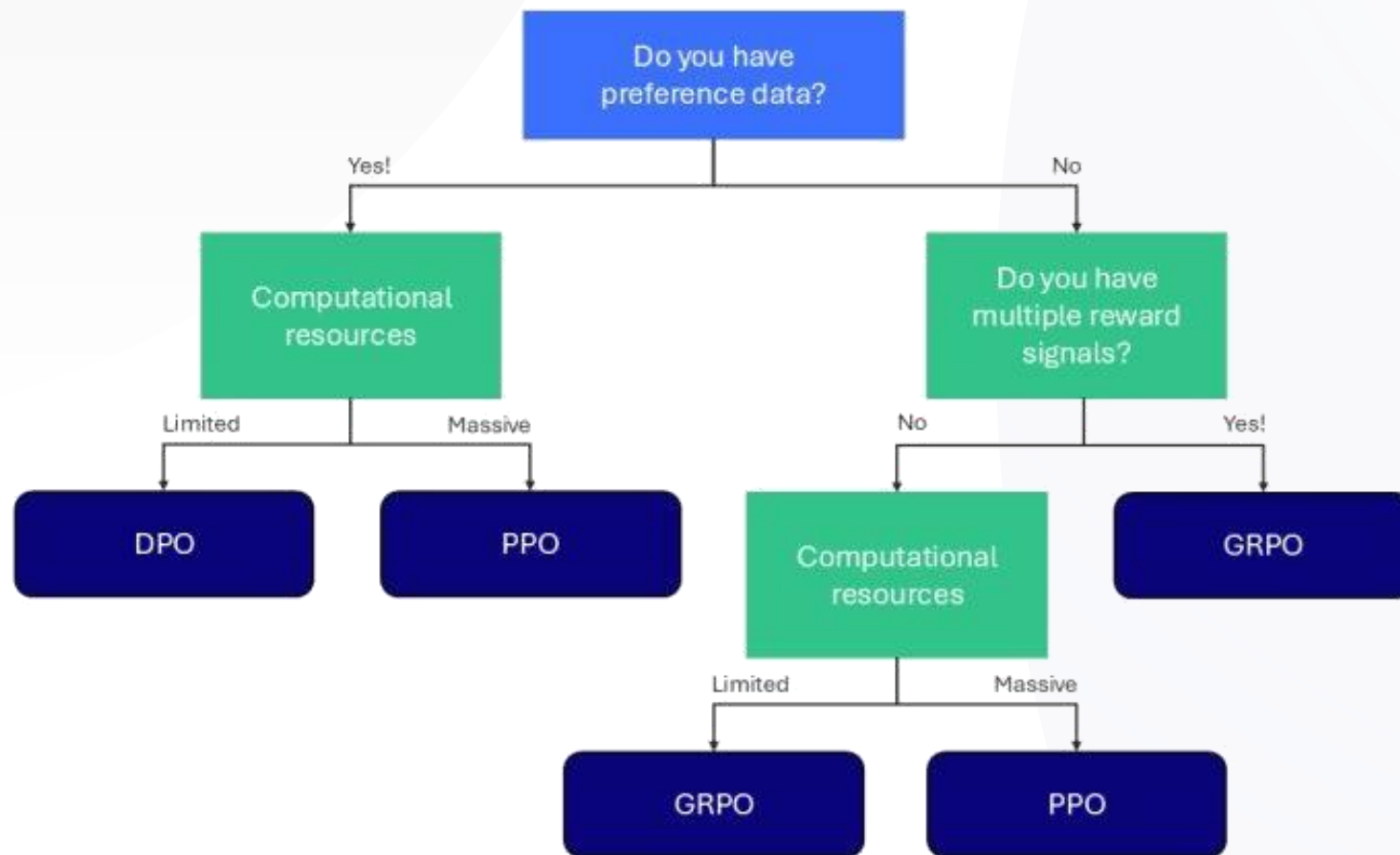
# Comparison of methods





RL Enhanced LLMs	Organization	# Params	RL Methods
Instruct-GPT (Ouyang et al., 2022)	 OpenAI	1.3B, 6B, 175B	RLHF, PPO
GPT-4 (OpenAI, 2023)	 OpenAI	-	RLHF, PPO, RBRM
Gemini (Team et al., 2023)		-	RLHF
InternLM2 (Cai et al., 2024)	 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory	1.8B, 7B, 20B	RLHF, PPO
Claude 3 (Anthropic, 2024)	<b>ANTHROPIC</b>	-	RLAIF
Reka (Team et al., 2024c)	 Reka	7B, 21B	RLHF, PPO
Zephyr (HuggingFaceH4, 2024)	 Argilla	141B-A39B	ORPO
Phi-3 (Abdin et al., 2024)	 Microsoft	3.8B, 7B, 14B	DPO
DeepSeek-V2 (Liu et al., 2024a)	 deepseek	236B-A21B	GRPO
ChatGLM (GLM et al., 2024)	 ZHIPU·AI	6B, 9B	ChatGLM-RLHF
Nemotron-4 340B (Adler et al., 2024)	 NVIDIA	340B	DPO, RPO
Llama 3 (Dubey et al., 2024)	 Meta	8B, 70B, 405B	DPO
Qwen2 (Yang et al., 2024a)	 Alibaba	(0.5-72)B, 57B-A14B	DPO
Gemma2 (Team et al., 2024b)		2B, 9B, 27B	RLHF
Starling-7B (Zhu et al., 2024)	 Berkeley UNIVERSITY OF CALIFORNIA	7B	RLAIF, PPO
Athene-70B (Nexusflow, 2024)	 Nexusflow	70B	RLHF
Hermes 3 (Teknium et al., 2024)	 NOUS RESEARCH	8B, 70B, 405B	DPO
o1 (OpenAI, 2024b)	 OpenAI	-	RL through CoT

# When to use DPO vs. PPO vs. GRPO?



**Thank you!**



## Bibliography and citations:

- Schulman, John, et al. "Proximal policy optimization algorithms." (2017).
- Lee, Harrison, et al. "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback (2023).
- Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in Neural Information Processing Systems 36 (2023).
- Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." (2024).
- Wang, Shuhe, et al. "Reinforcement Learning Enhanced LLMs: A Survey



## Image references:

- Image 1: Cao, Yuji, et al. "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods." IEEE Transactions on Neural Networks and Learning Systems (2024).
- Image 2,3: [Play the Game](#) and [N-Step Lookahead](#)
- Image 4,5,6,7,8: [LLM Training: RLHF and Its Alternatives](#)
- Dataset: [Skywork/Skywork-Reward-Preference-80K-v0.2 · Datasets at Hugging Face](#)
- Image 9: Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).
- Image 10,11,12,13,14: Lee, Harrison, et al. "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback (2023).
- Image 15: Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in Neural Information Processing Systems 36 (2023).
- Image 16: Wang, Shuhe, et al. "Reinforcement Learning Enhanced LLMs: A Survey

