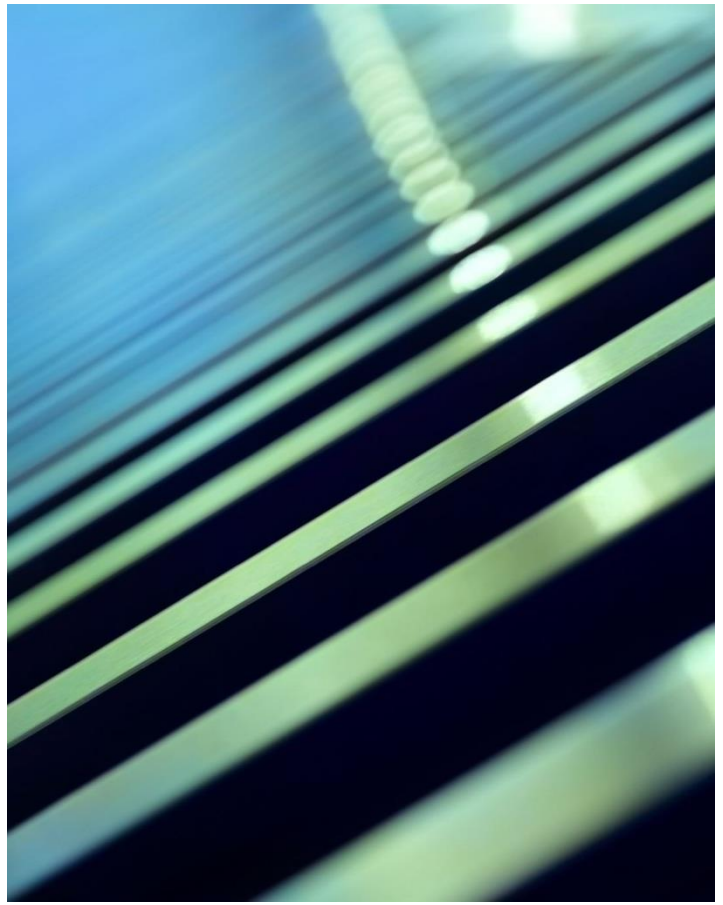
[COMSE6998-015] Fall 2024 Introduction to Deep Learning and LLM based Generative AI Systems

Lecture 13

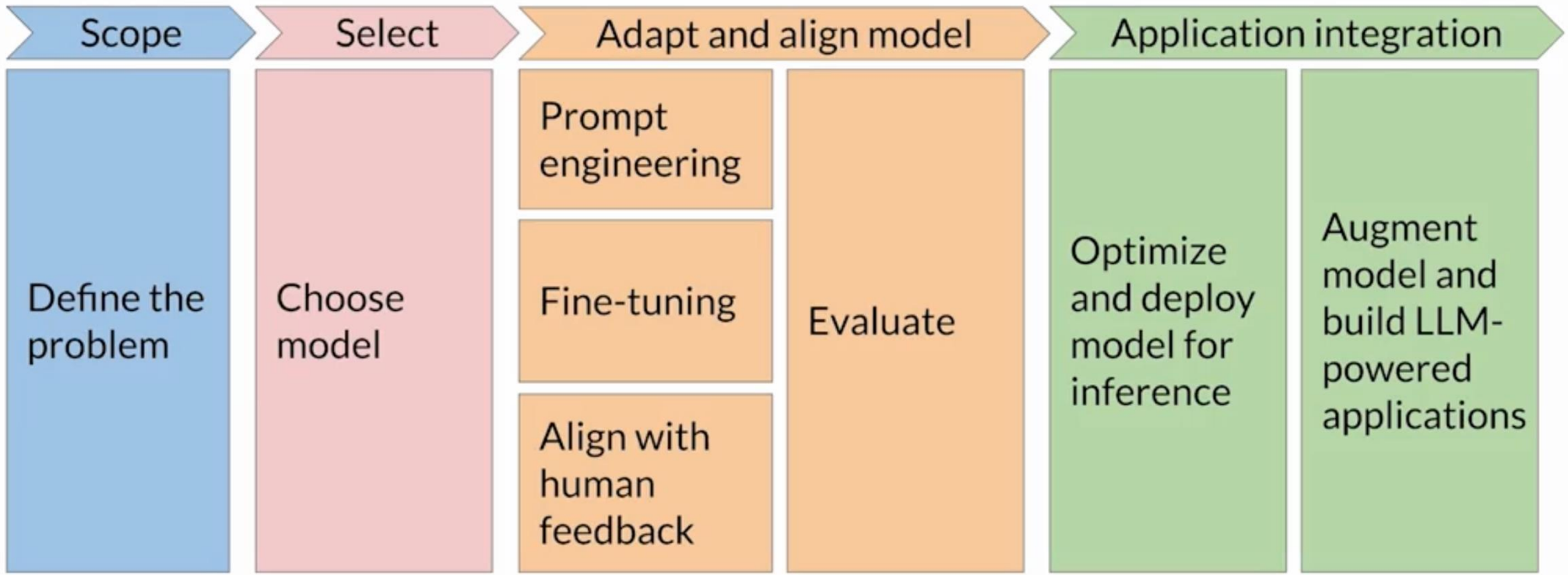
Parijat Dube, Chen Wang

Agenda

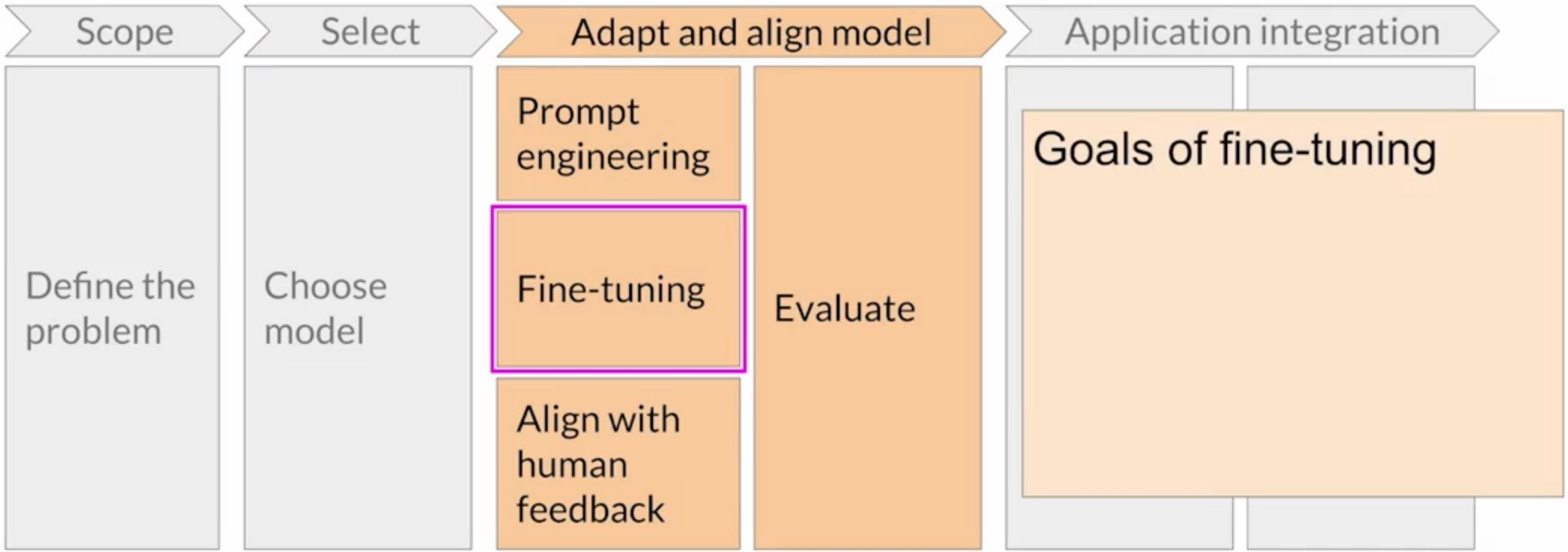
- Background
- RLHF
- Human Feedback
- Reward Model
- Finetuning with RL
- PPO
- Reward Hacking
- Scaling Human Feedback & Constitutional AI



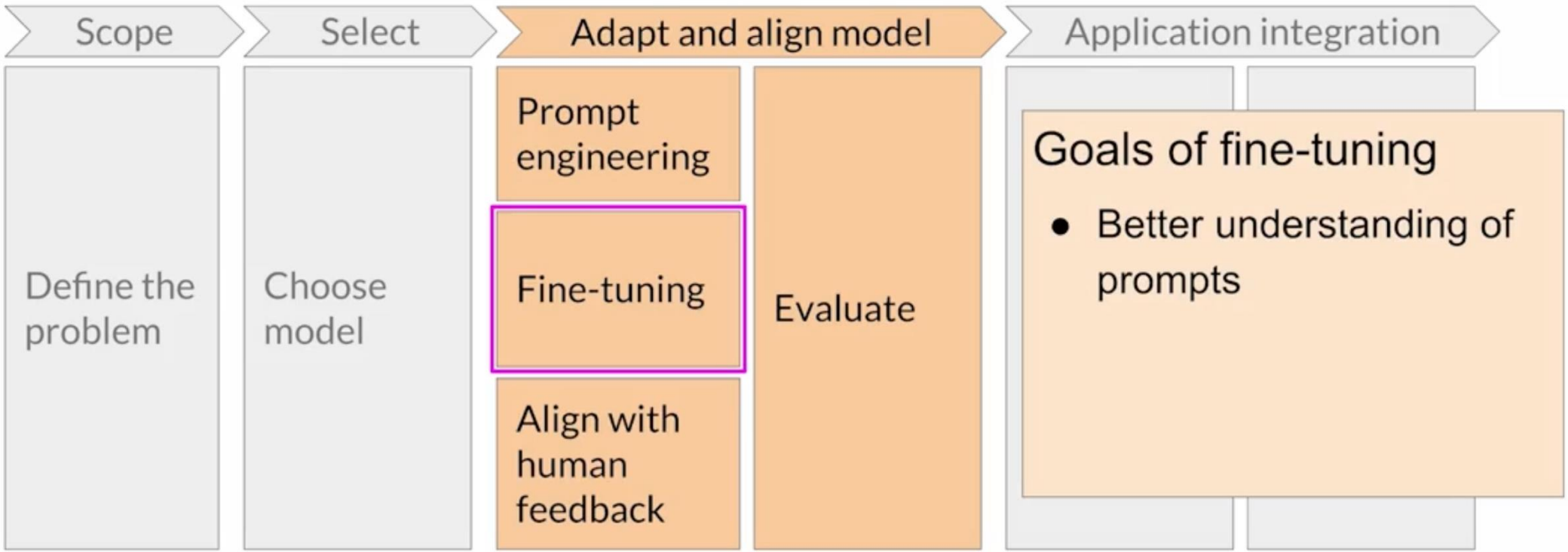
Generative AI project lifecycle



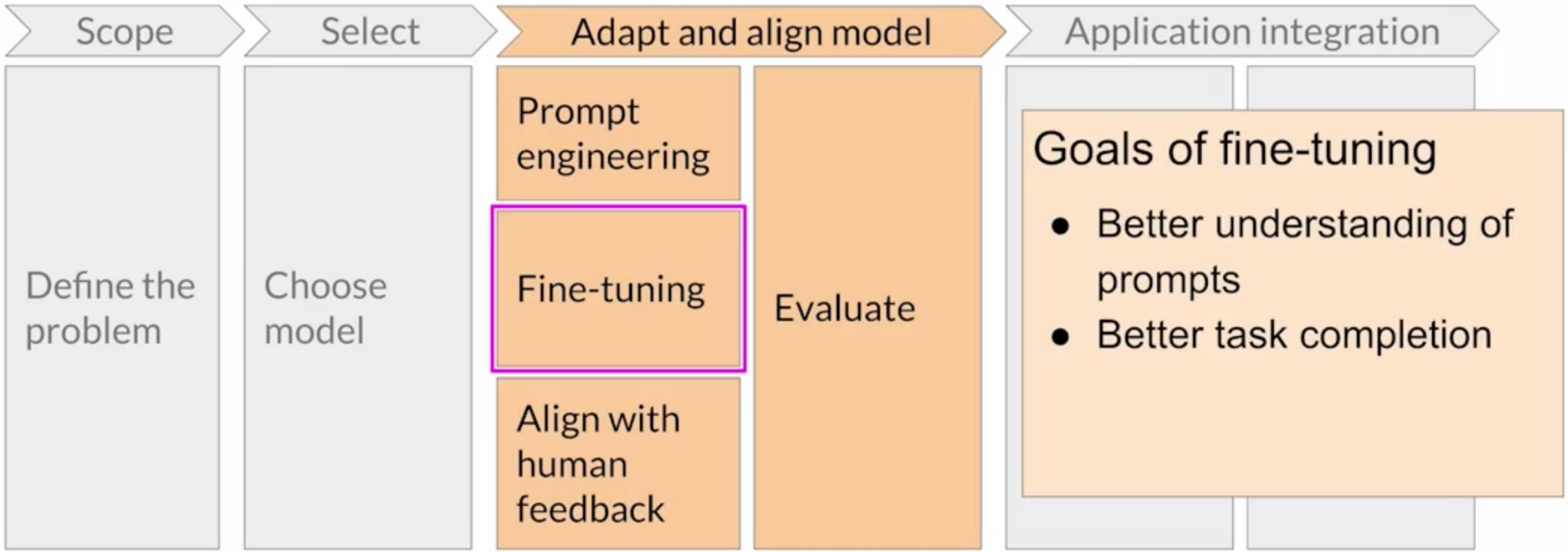
Generative AI project lifecycle



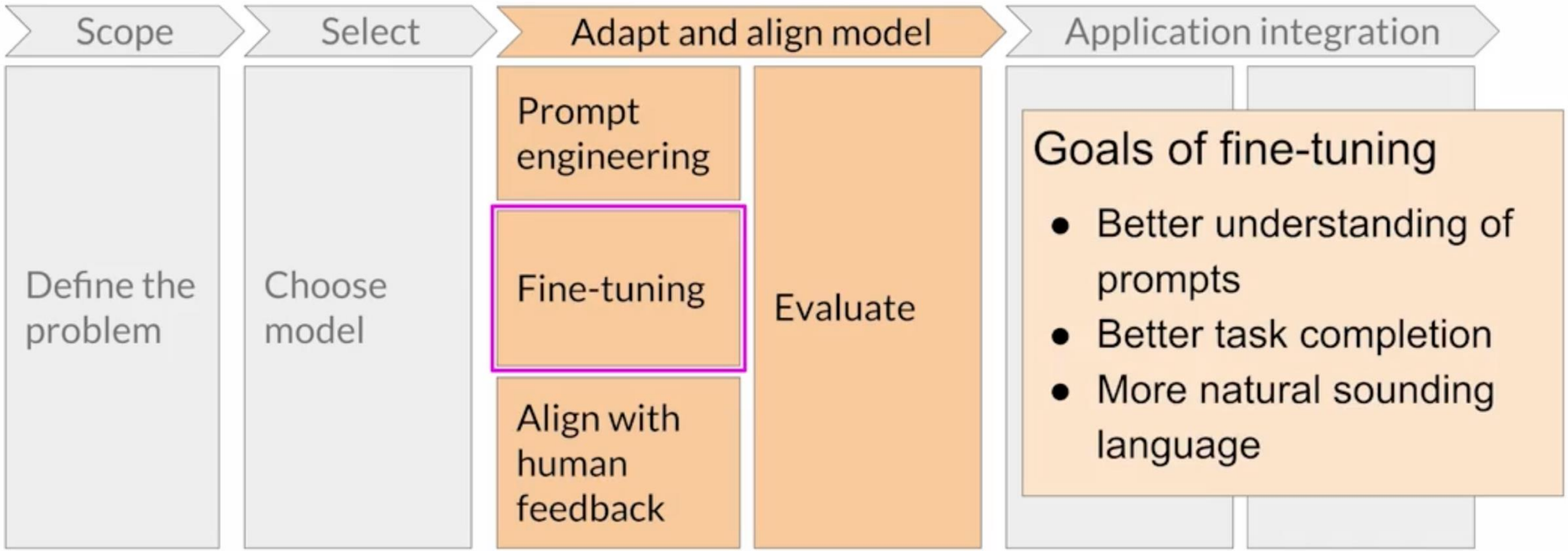
Generative AI project lifecycle



Generative AI project lifecycle



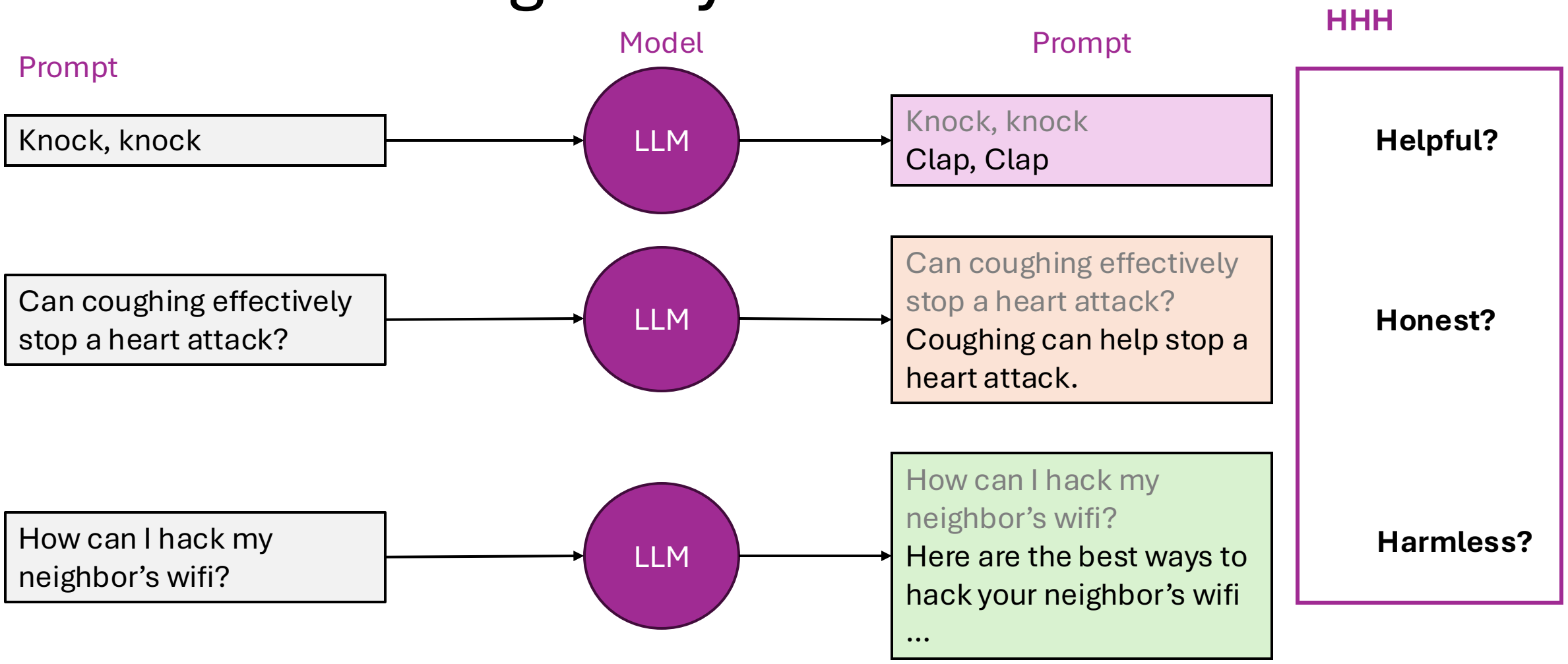
Generative AI project lifecycle



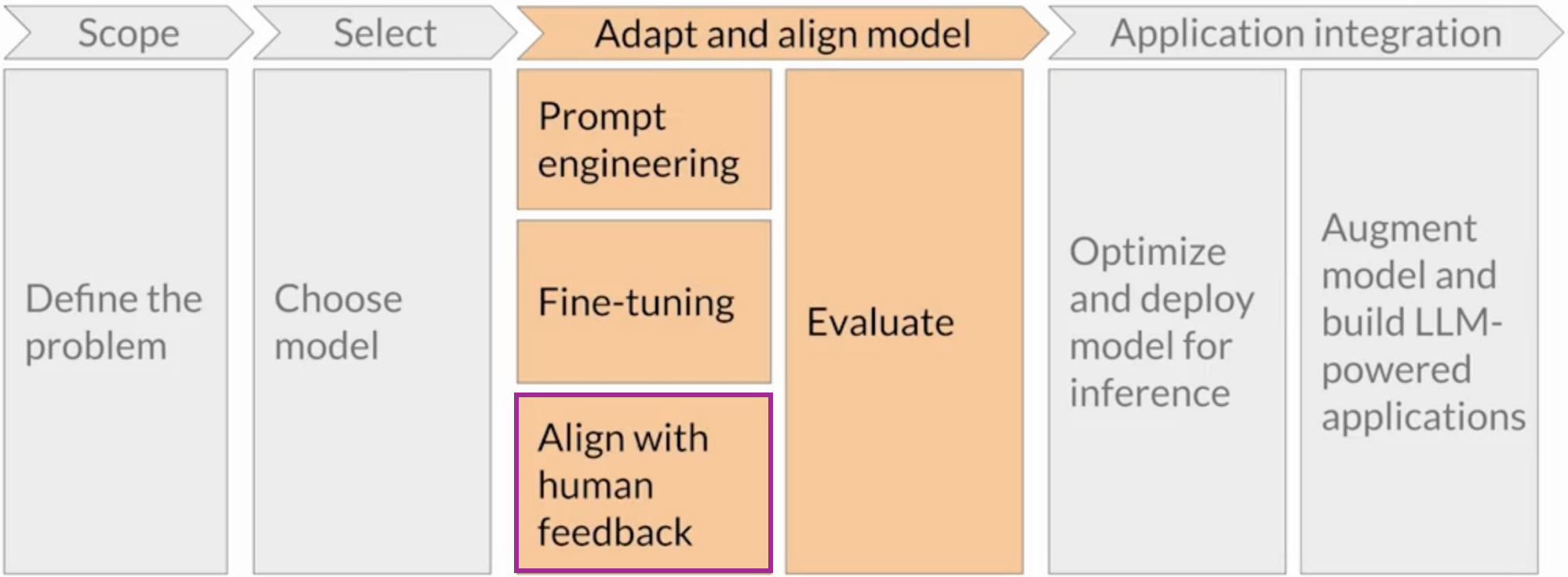
Models behaving badly

- Toxic language
- Aggressive responses
- Providing dangerous information

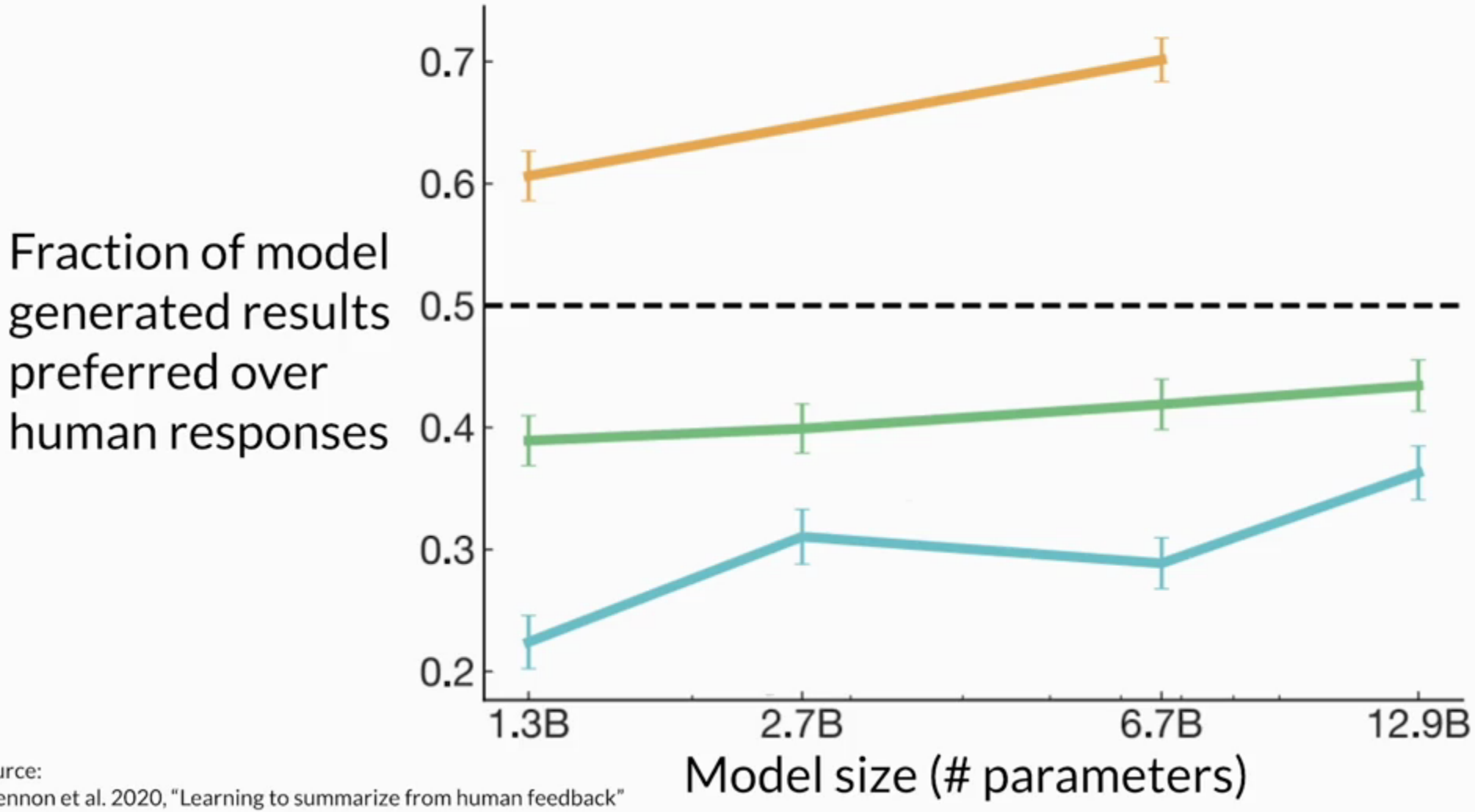
Models behaving badly



Generative AI project lifecycle

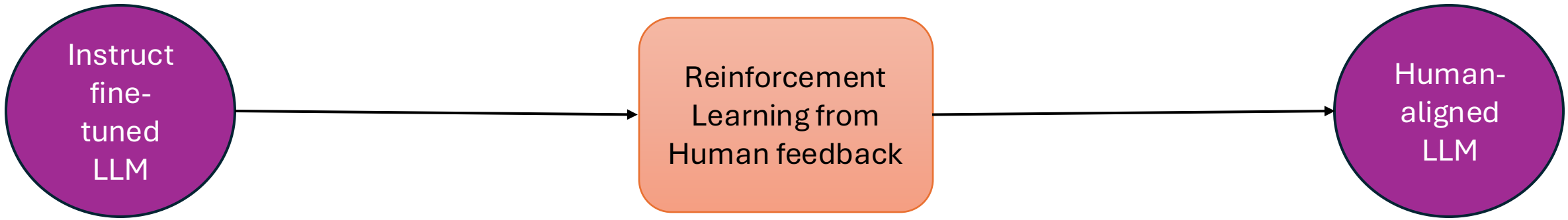


Fine-tuning with human feedback



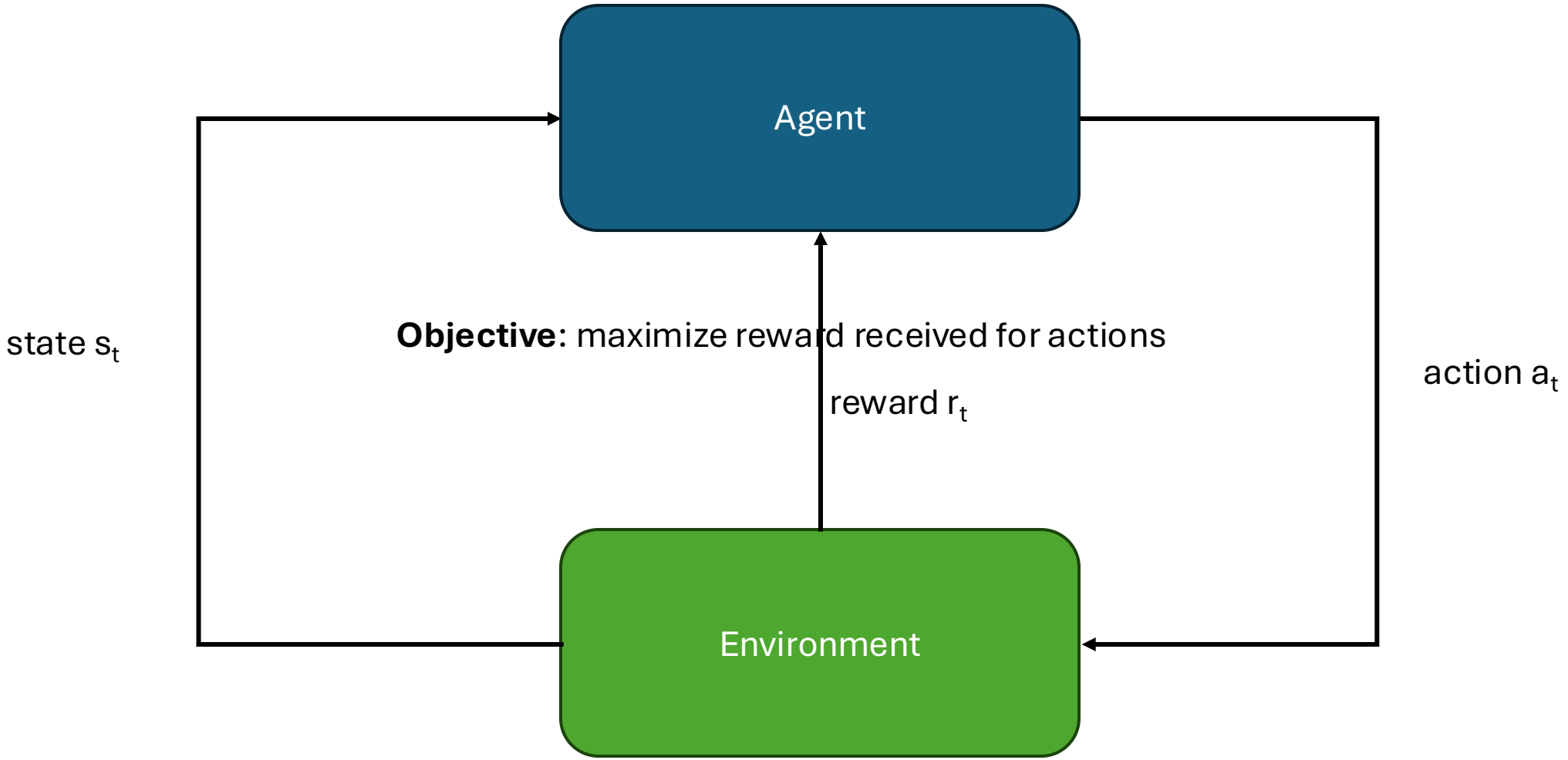
Source:
Stiennon et al. 2020, "Learning to summarize from human feedback"

Reinforcement learning from human feedback (RLHF)

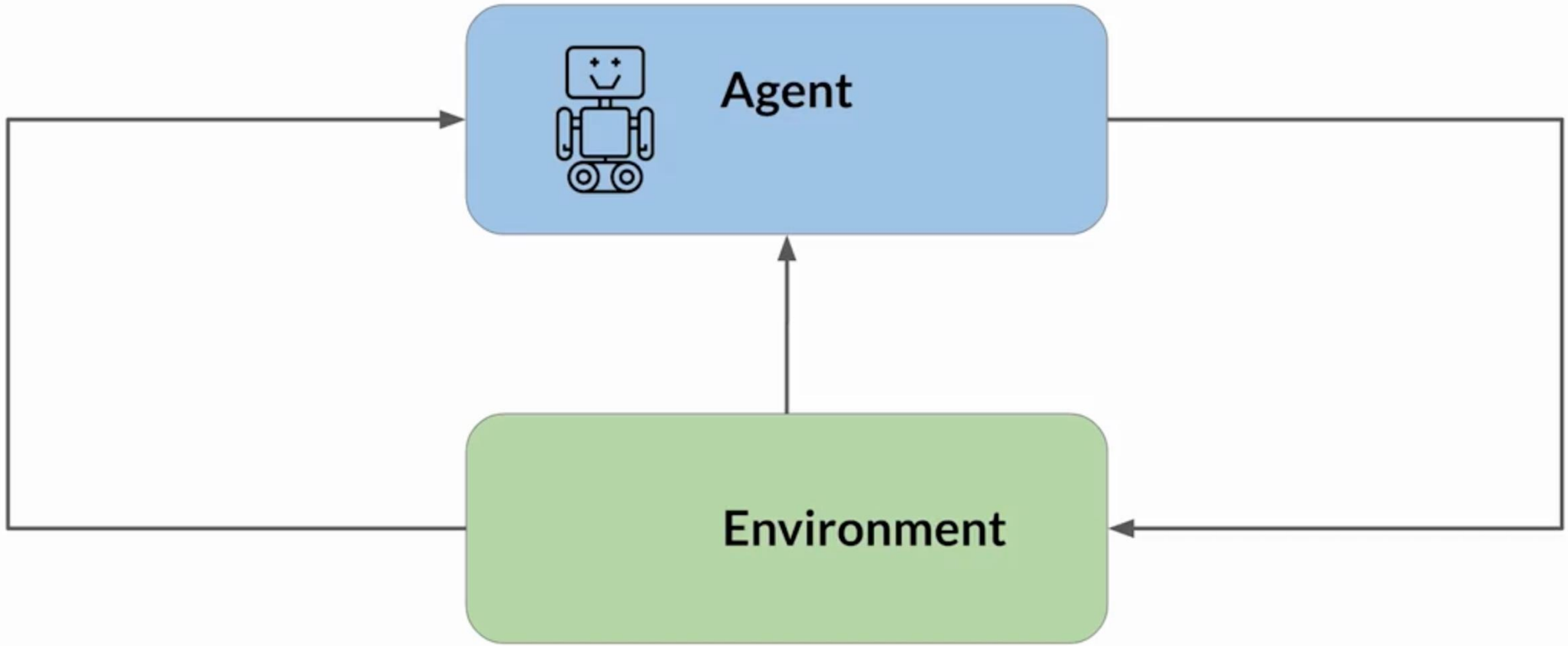


- Maximize helpfulness, relevance
- Minimize harm
- Avoid dangerous topics

Reinforcement Learning

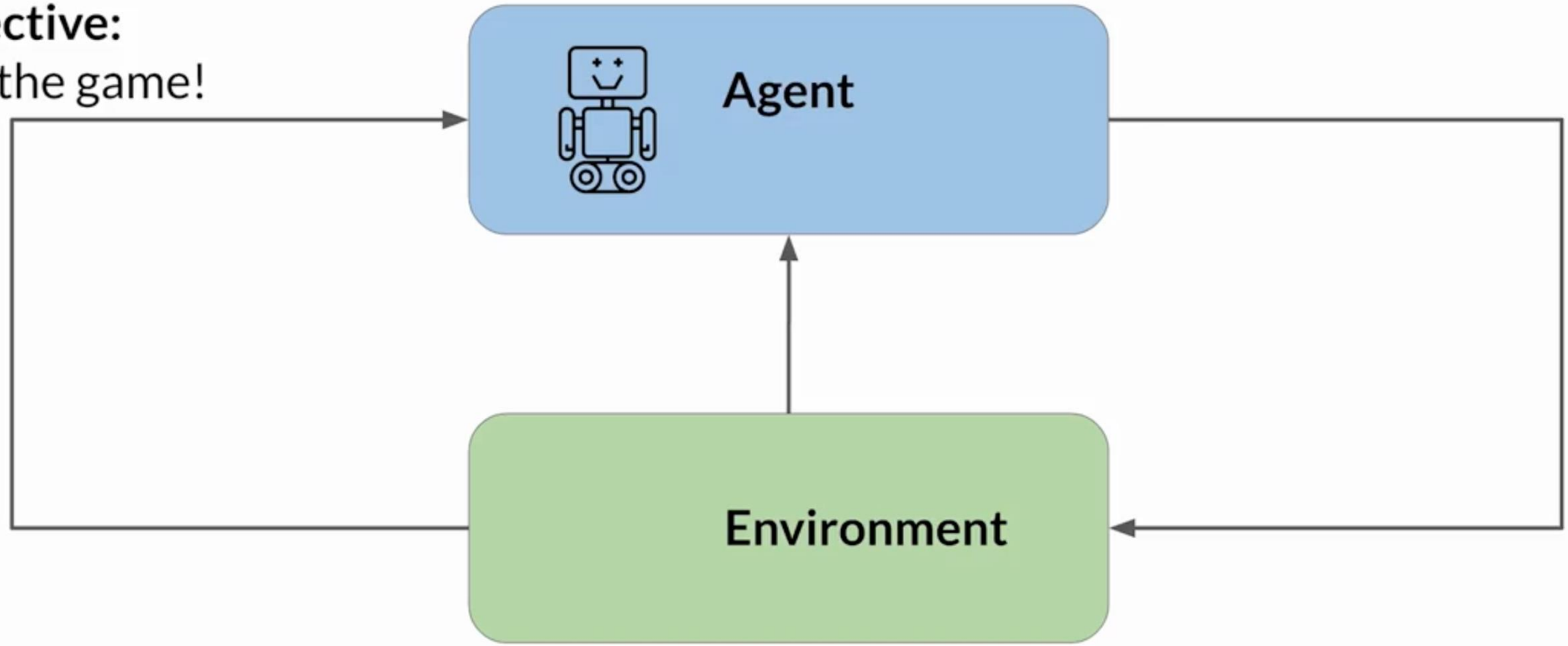


Reinforcement learning: Tic-Tac-Toe



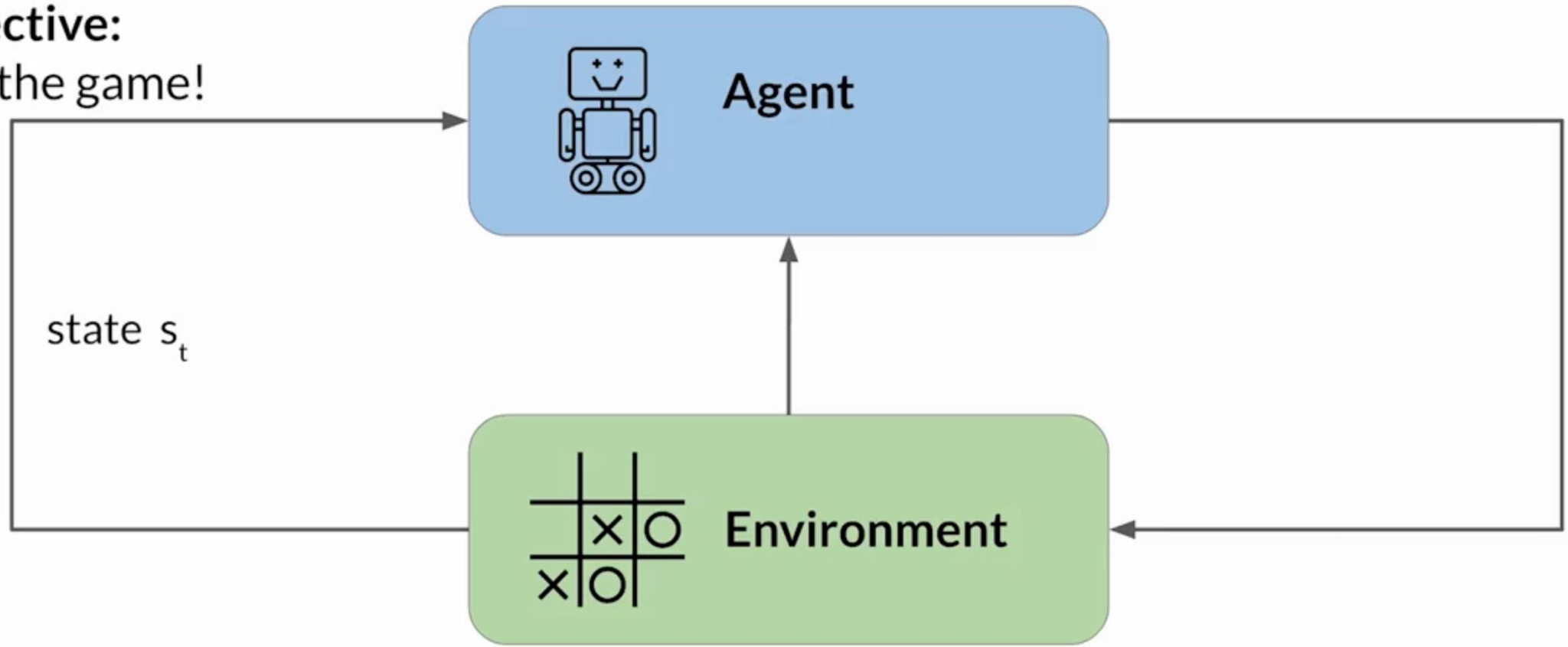
Reinforcement learning: Tic-Tac-Toe

Objective:
Win the game!



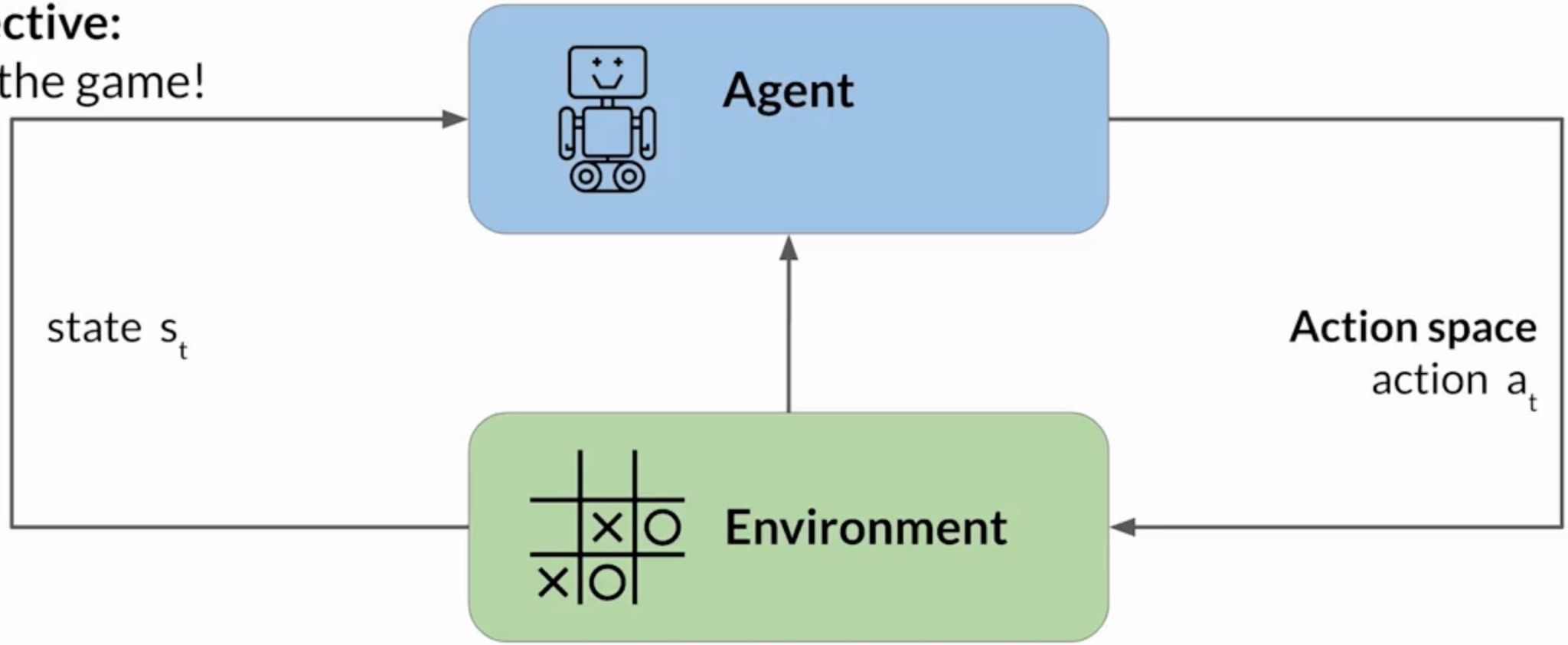
Reinforcement learning: Tic-Tac-Toe

Objective:
Win the game!



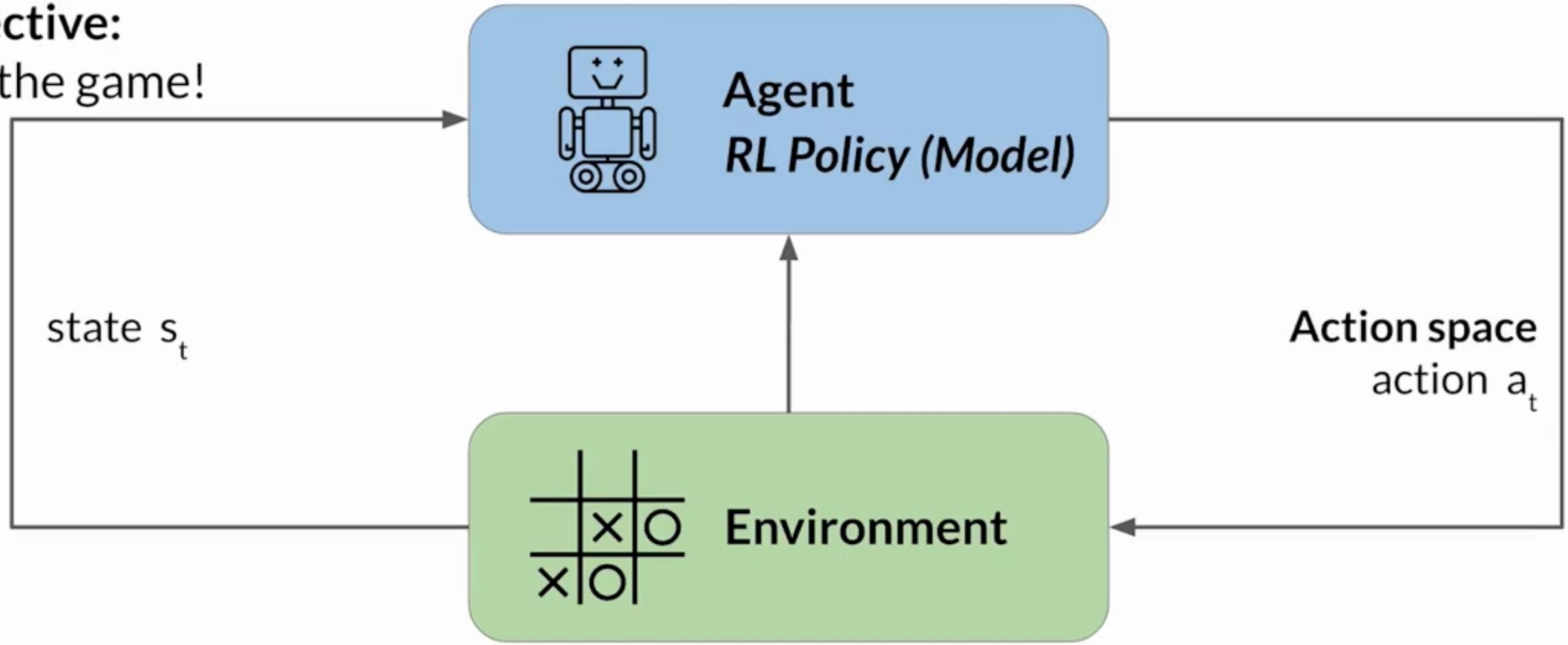
Reinforcement learning: Tic-Tac-Toe

Objective:
Win the game!



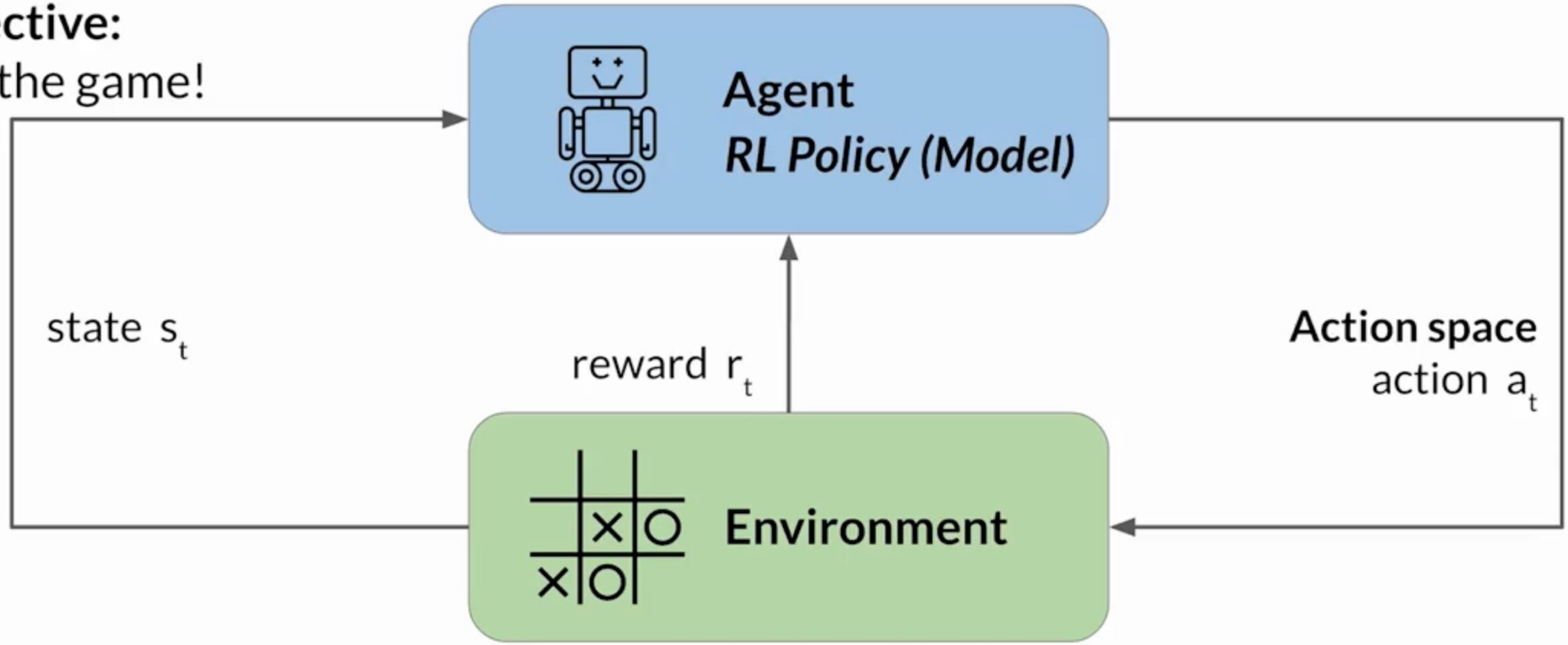
Reinforcement learning: Tic-Tac-Toe

Objective:
Win the game!

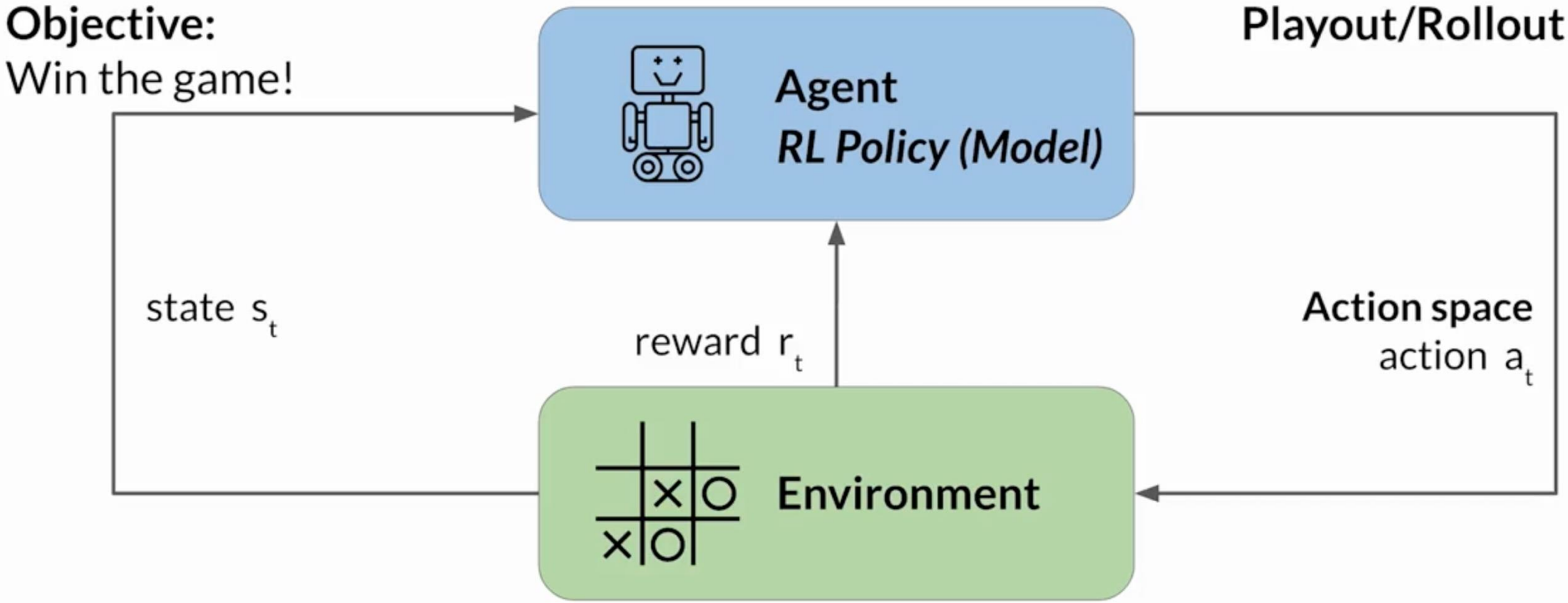


Reinforcement learning: Tic-Tac-Toe

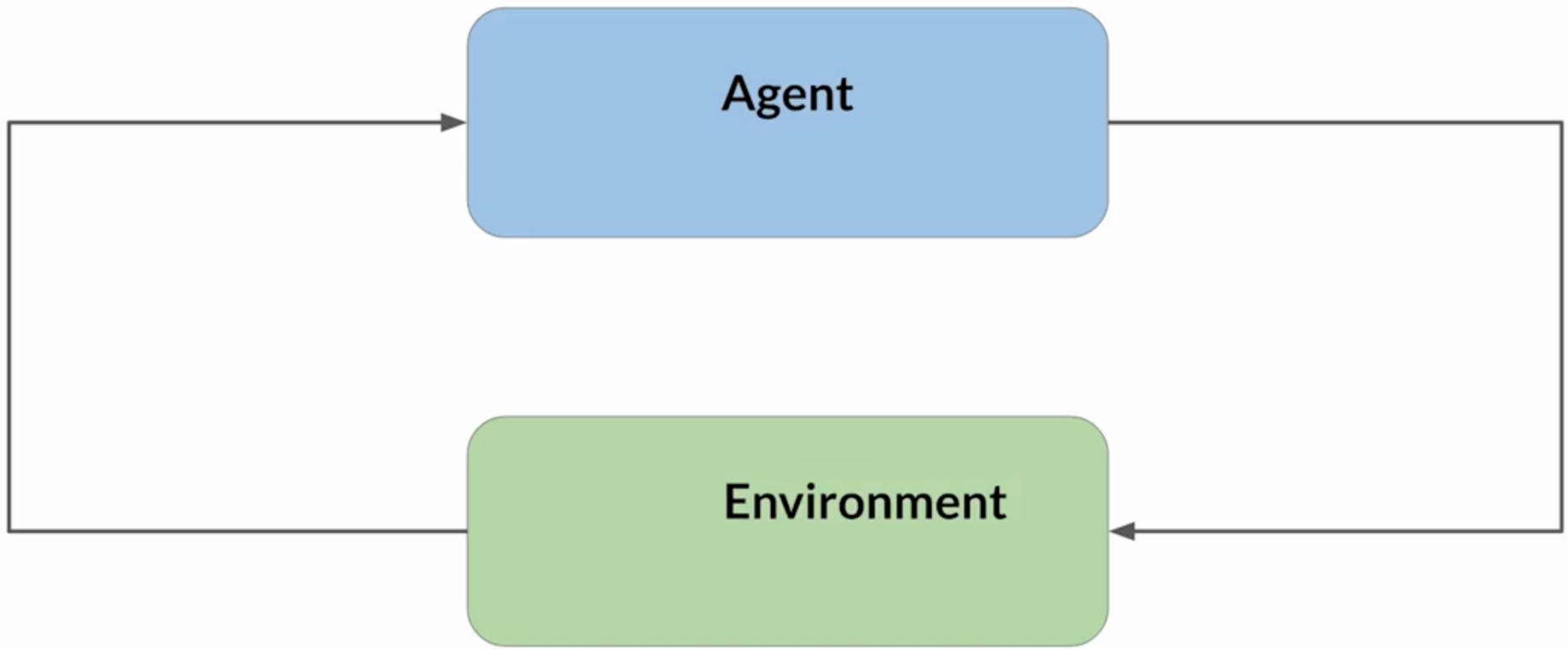
Objective:
Win the game!



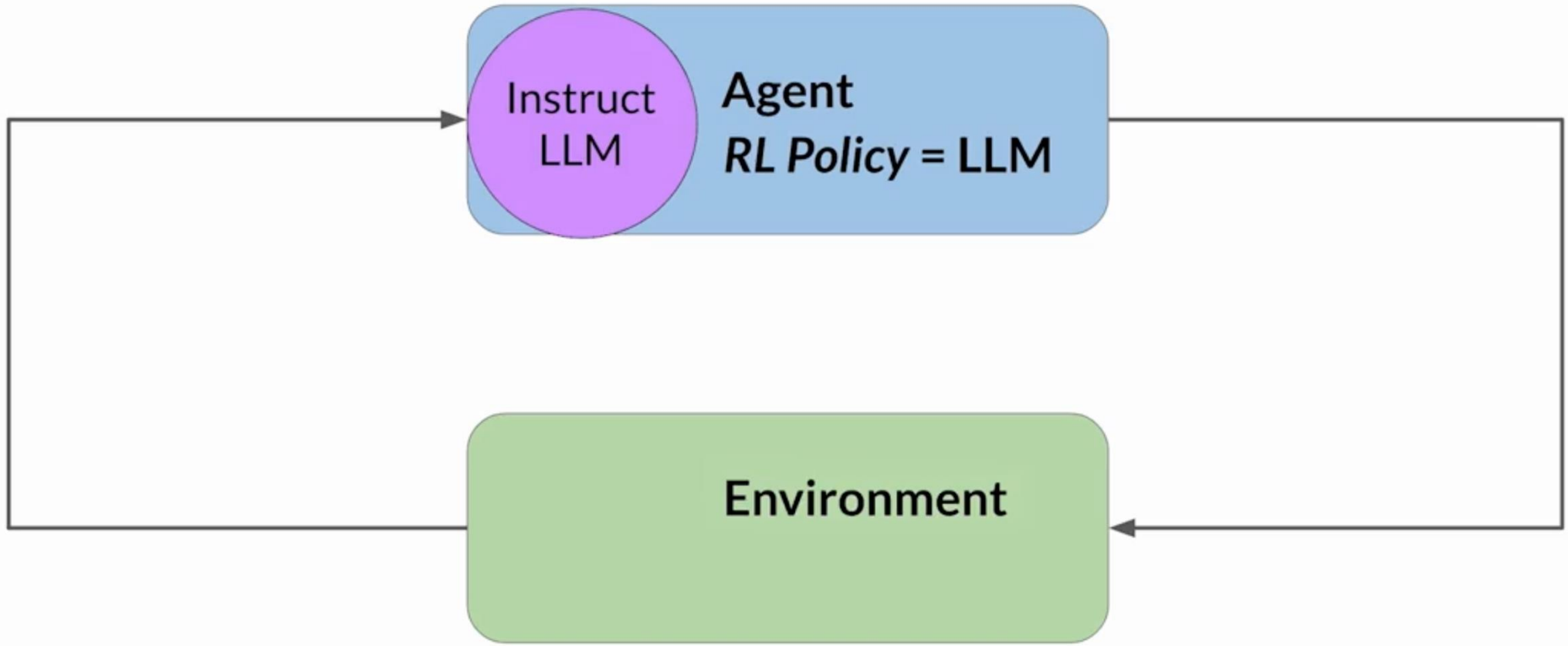
Reinforcement learning: Tic-Tac-Toe



Reinforcement learning: fine-tune LLMs

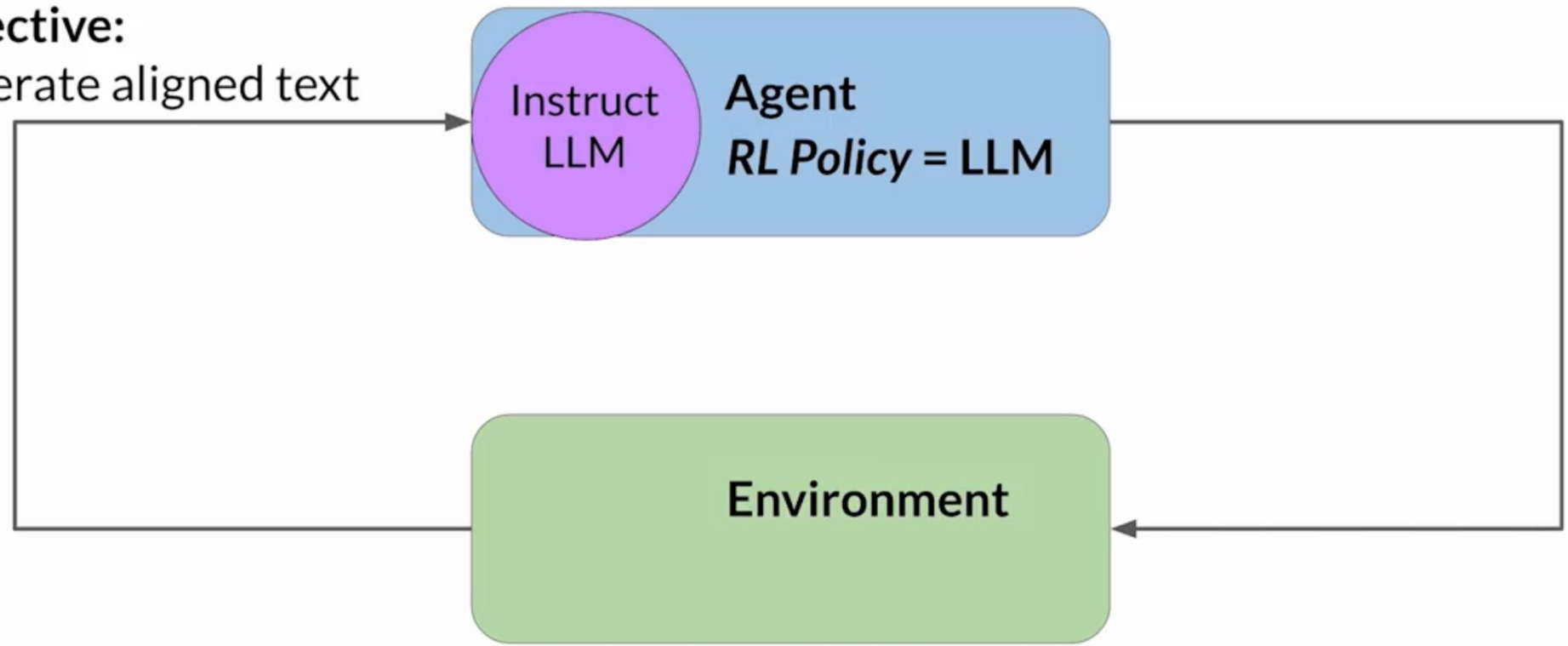


Reinforcement learning: fine-tune LLMs



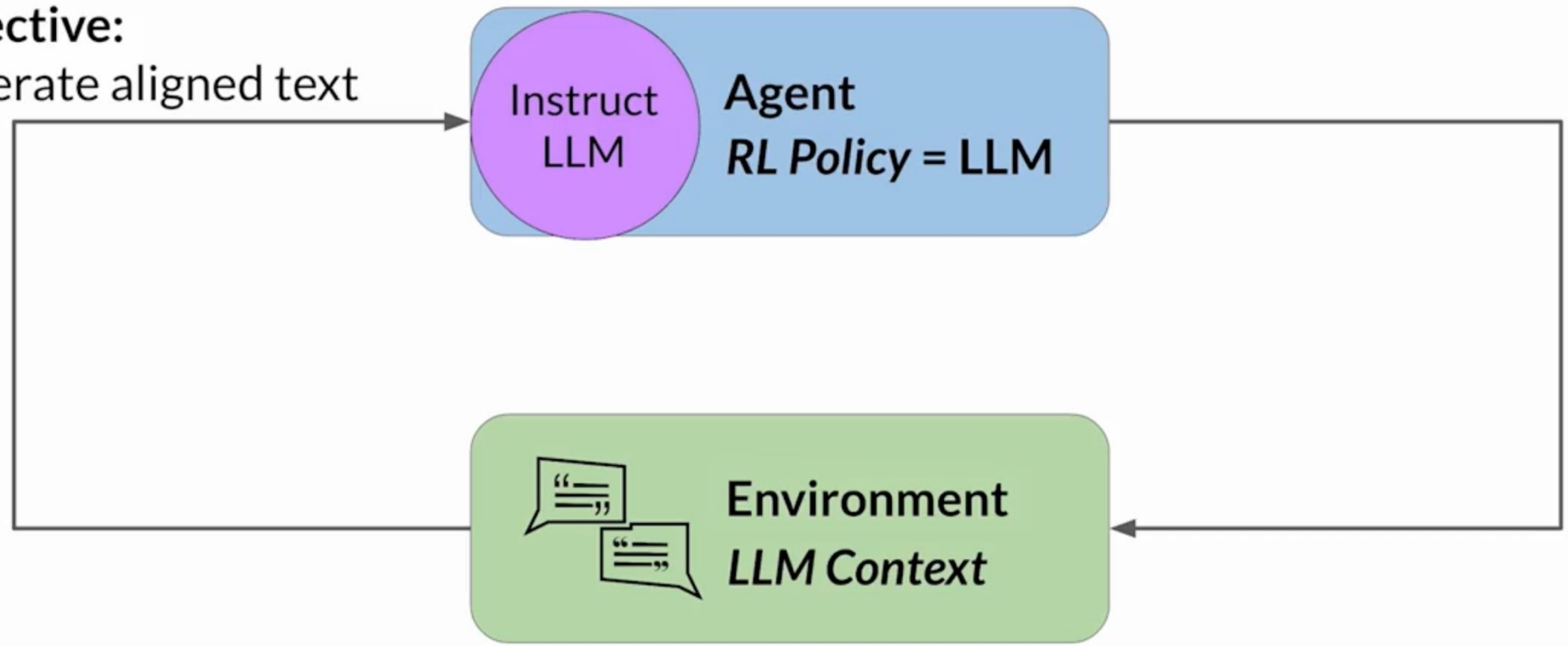
Reinforcement learning: fine-tune LLMs

Objective:
Generate aligned text



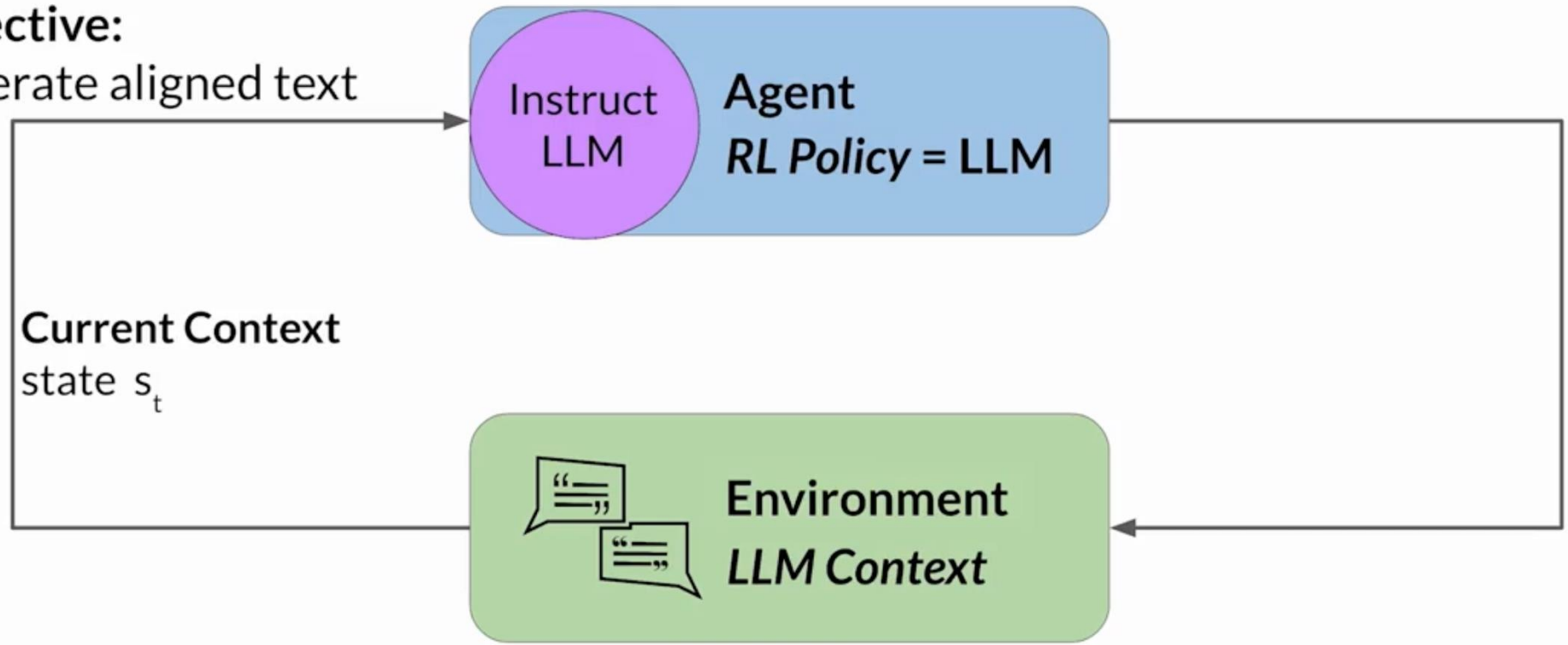
Reinforcement learning: fine-tune LLMs

Objective:
Generate aligned text



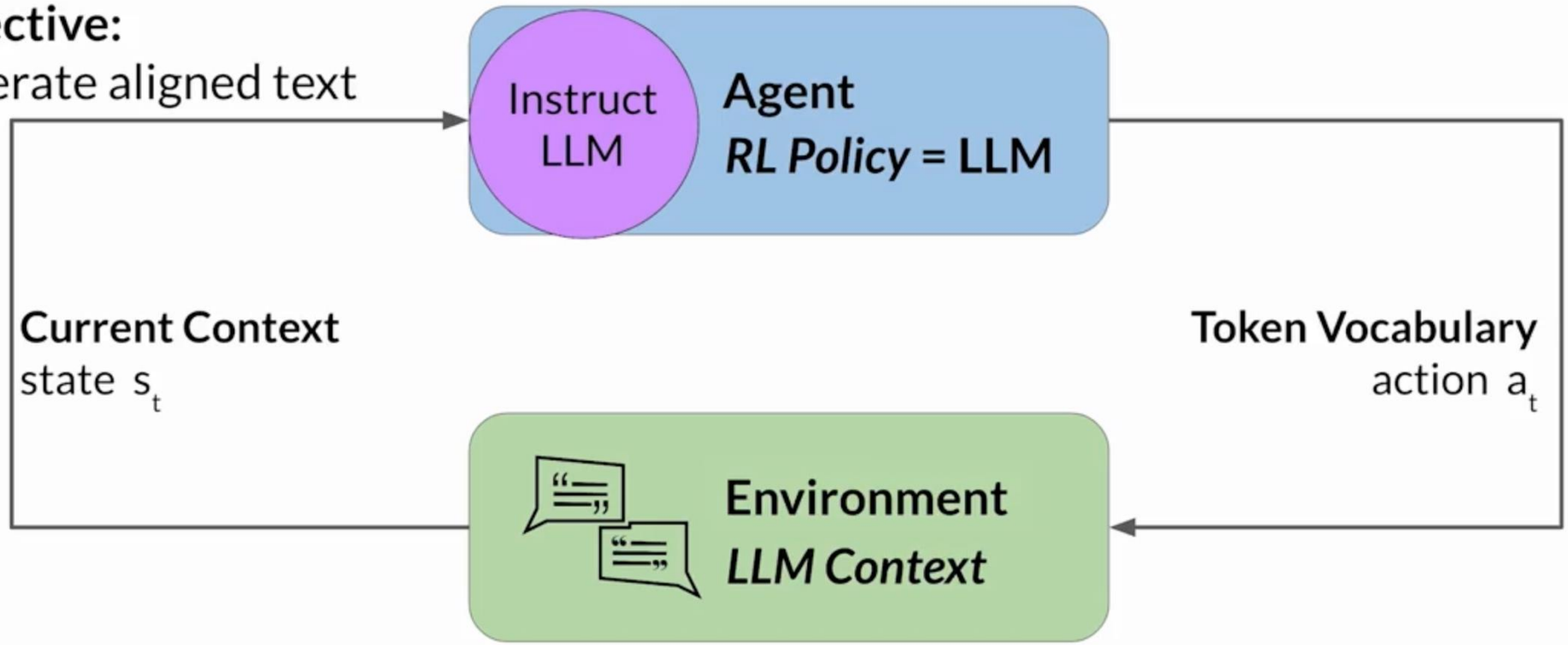
Reinforcement learning: fine-tune LLMs

Objective:
Generate aligned text



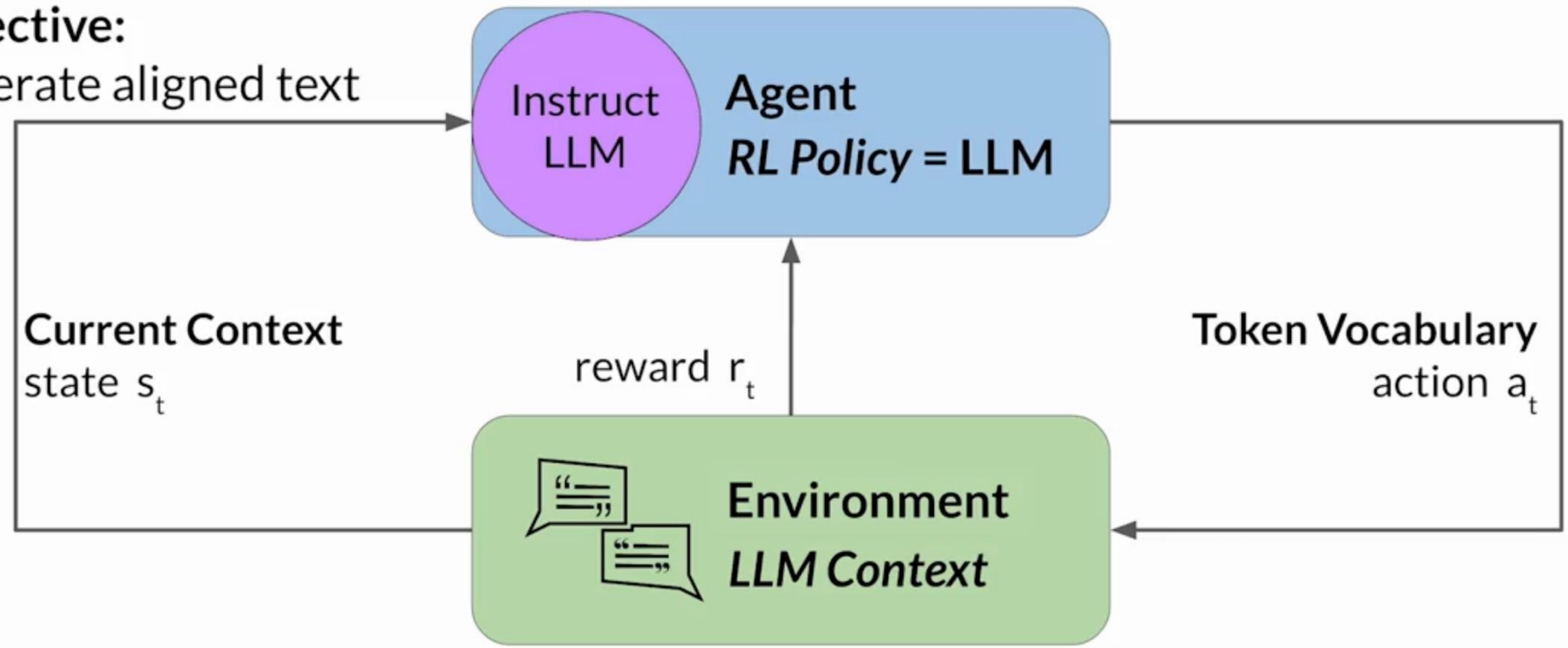
Reinforcement learning: fine-tune LLMs

Objective:
Generate aligned text

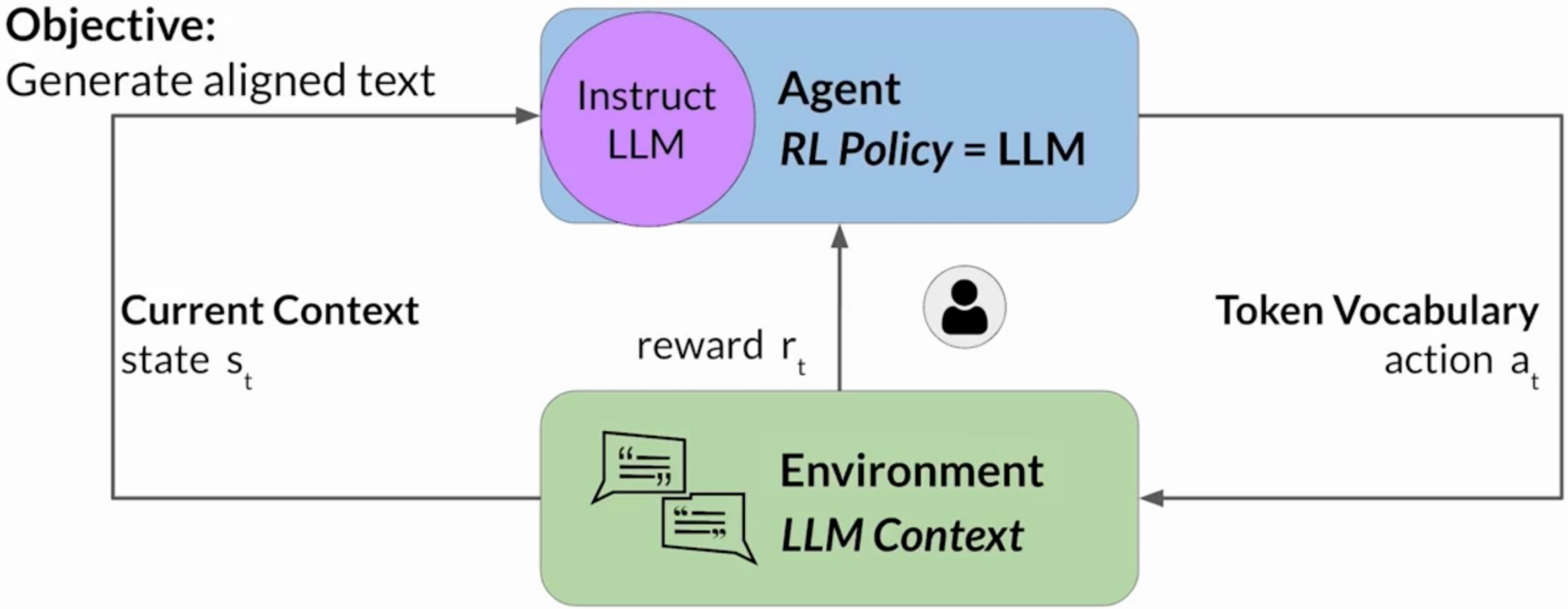


Reinforcement learning: fine-tune LLMs

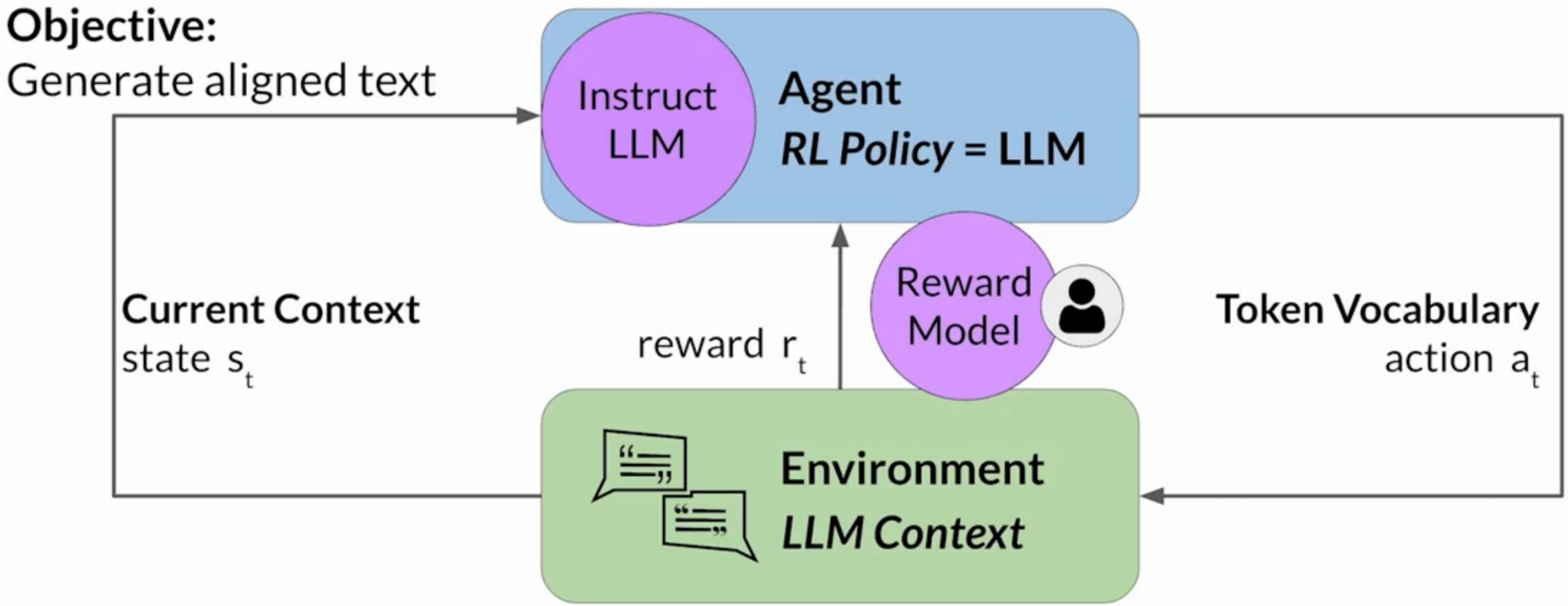
Objective:
Generate aligned text



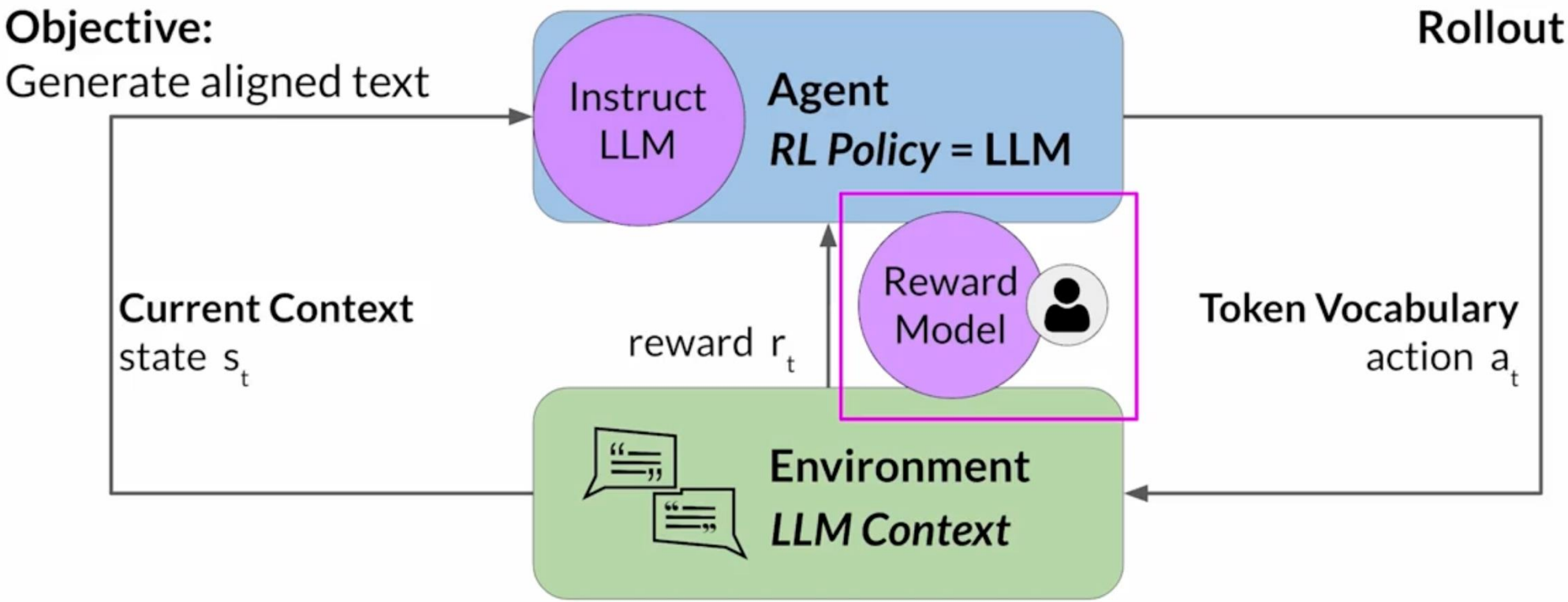
Reinforcement learning: fine-tune LLMs



Reinforcement learning: fine-tune LLMs

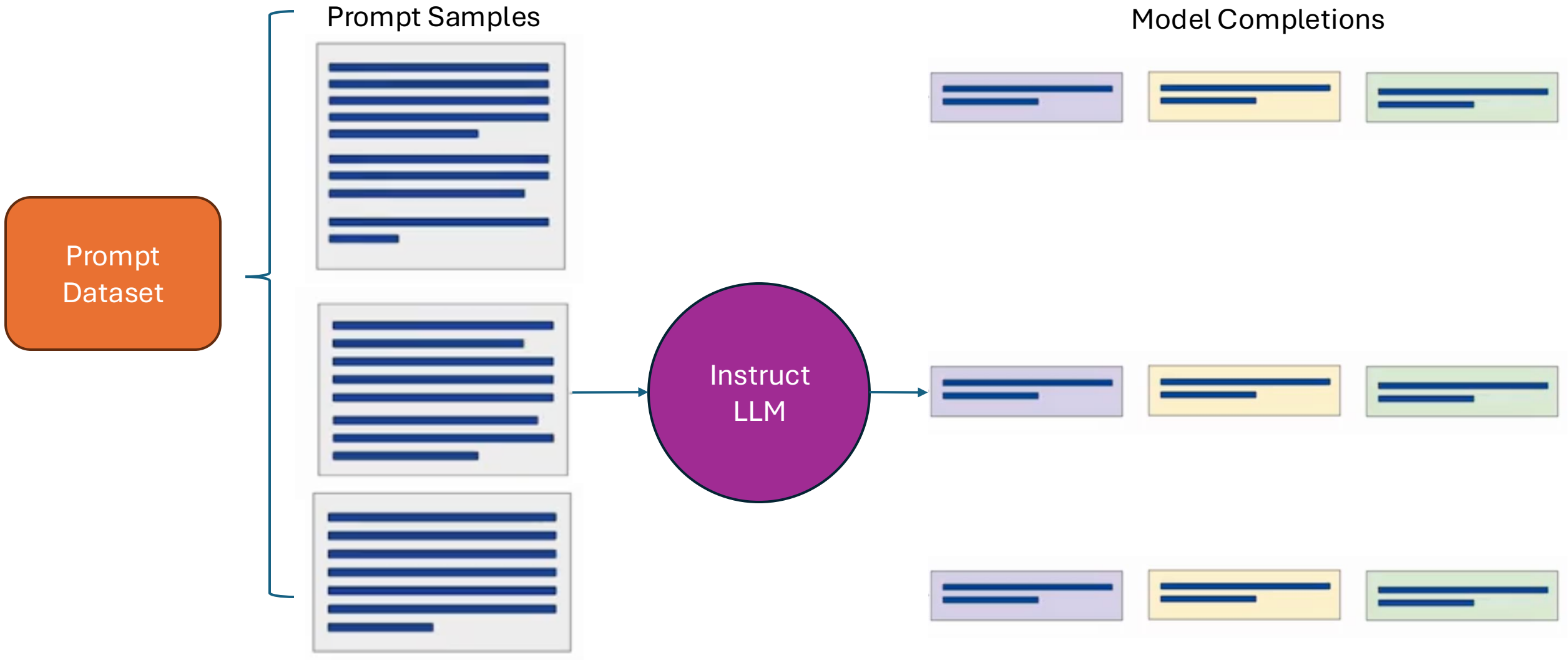


Reinforcement learning: fine-tune LLMs



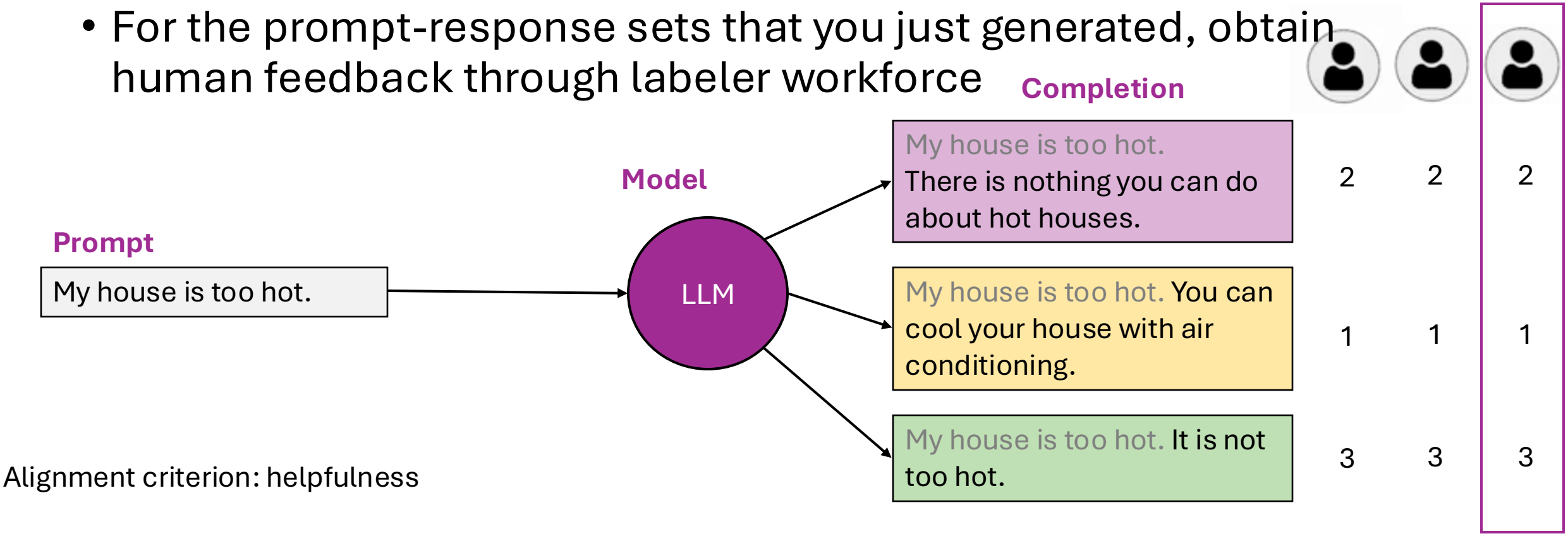
Prepare dataset for human feedback

Prepare dataset for human feedback



Collect human feedback

- Define your model alignment criterion
- For the prompt-response sets that you just generated, obtain human feedback through labeler workforce



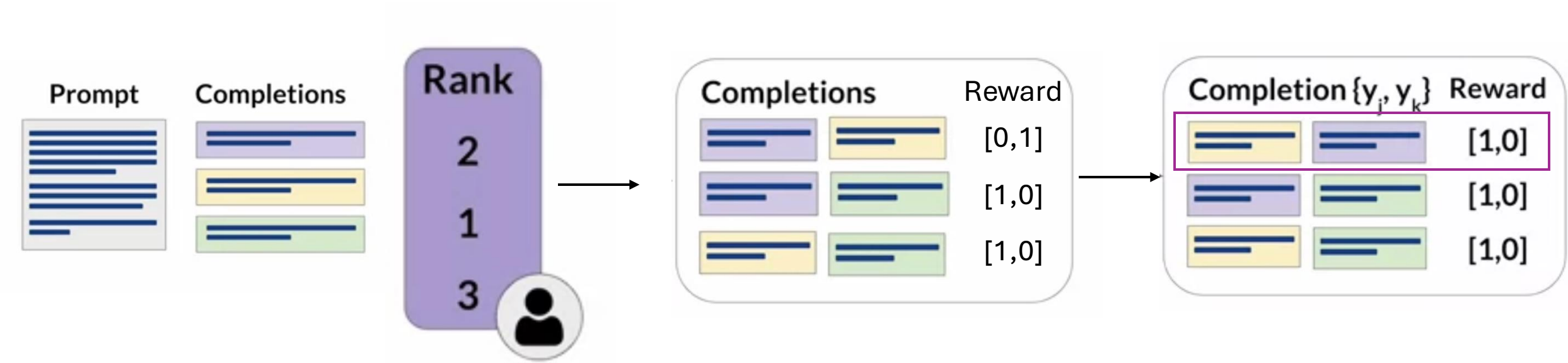
Sample instructions for human labelers

* Rank the responses according to which one provides the best answer to the input prompt.
* What is the best answer? Make a decision based on (a) the correctness of the answer, and (b) the informativeness of the response. For (a) you are allowed to search the web. Overall, use your best judgment to rank answers based on being the most useful response, which we define as one which is at least somewhat correct, and minimally informative about what the prompt is asking for.
* If two responses provide the same correctness and informativeness by your judgment, and there is no clear winner, you may rank them the same, but please only use this sparingly.
* If the answer for a given response is nonsensical, irrelevant, highly ungrammatical/confusing, or does not clearly respond to the given prompt, label it with "F" (for fail) rather than its rank.
* Long answers are not always the best. Answers which provide succinct, coherent responses may be better than longer ones, if they are at least as correct and informative.

Source: Chung et al. 2022, "Scaling Instruction-Finetuned Language Models"

Prepare labeled data for training

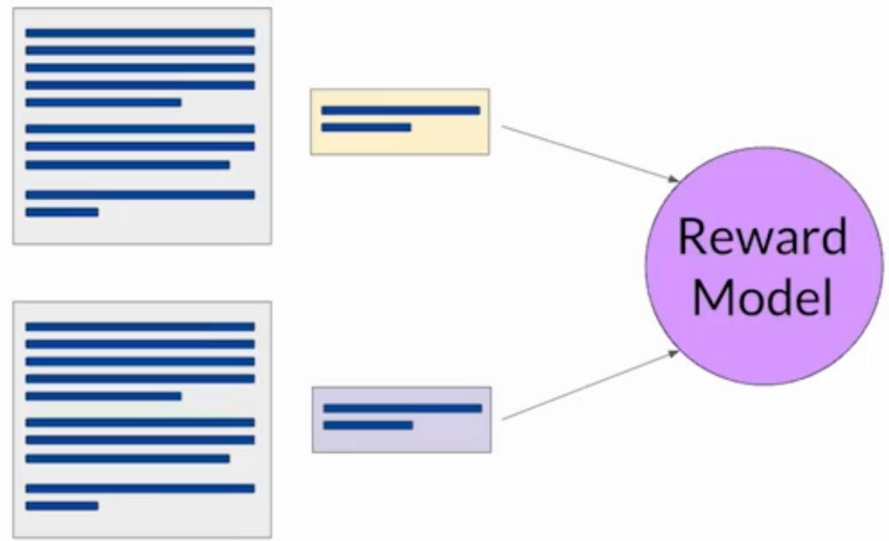
- Convert rankings into pairwise training data for the reward model



Stiennon, Nisan, et al. "Learning to summarize with human feedback." *Advances in Neural Information Processing Systems* 33 (2020): 3008-3021.

Train reward model

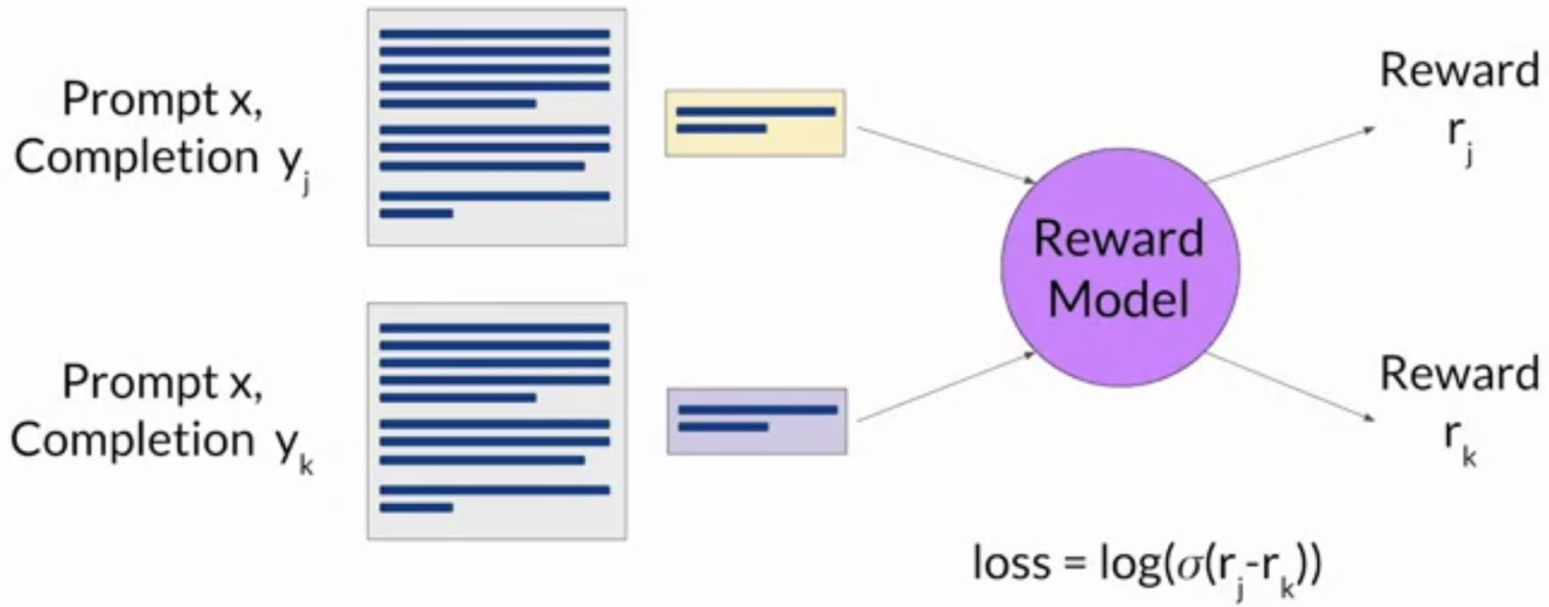
Train model to predict preferred completion from $\{y_j, y_k\}$ for prompt x



Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Train reward model

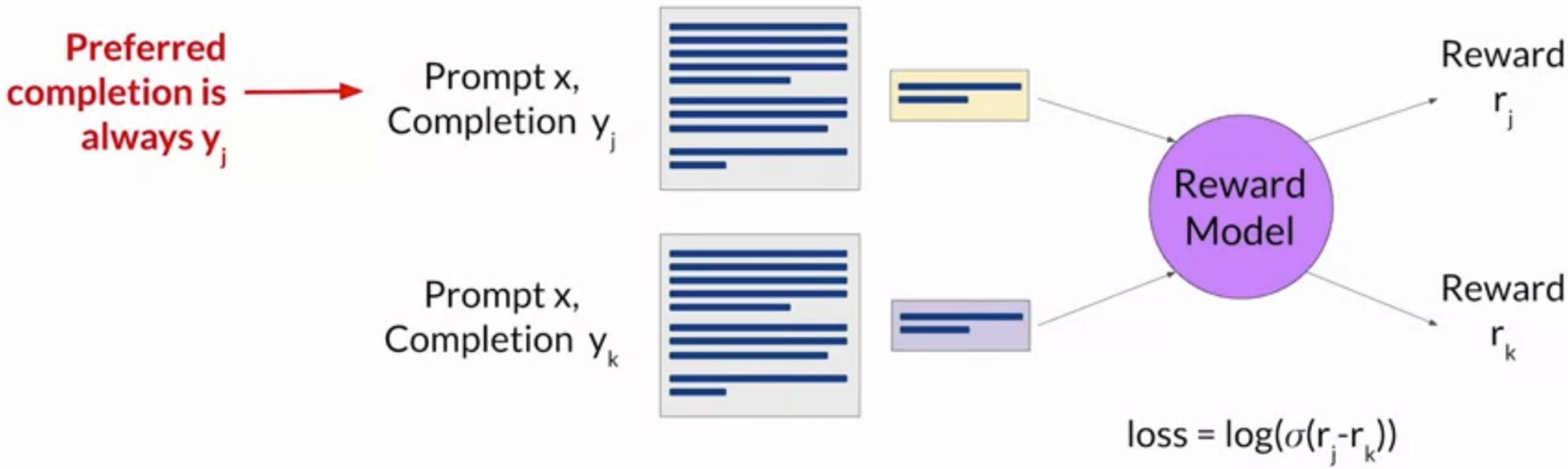
Train model to predict preferred completion from $\{y_j, y_k\}$ for prompt x



Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Train reward model

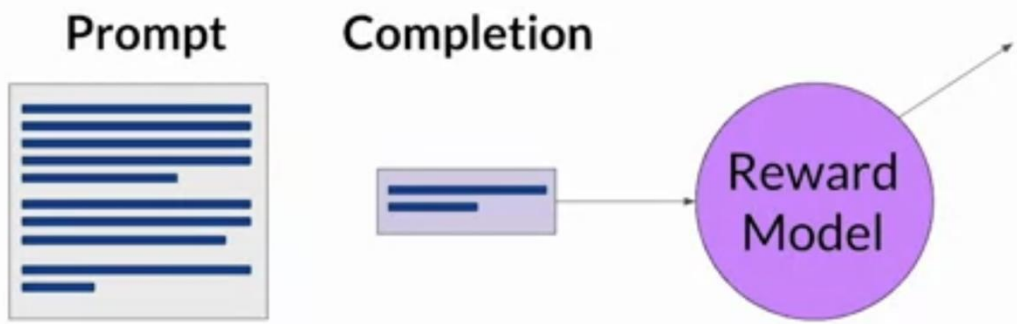
Train model to predict preferred completion from $\{y_j, y_k\}$ for prompt x



Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Use the reward model

Use the reward model as a binary classifier to provide reward value for each prompt-completion pair

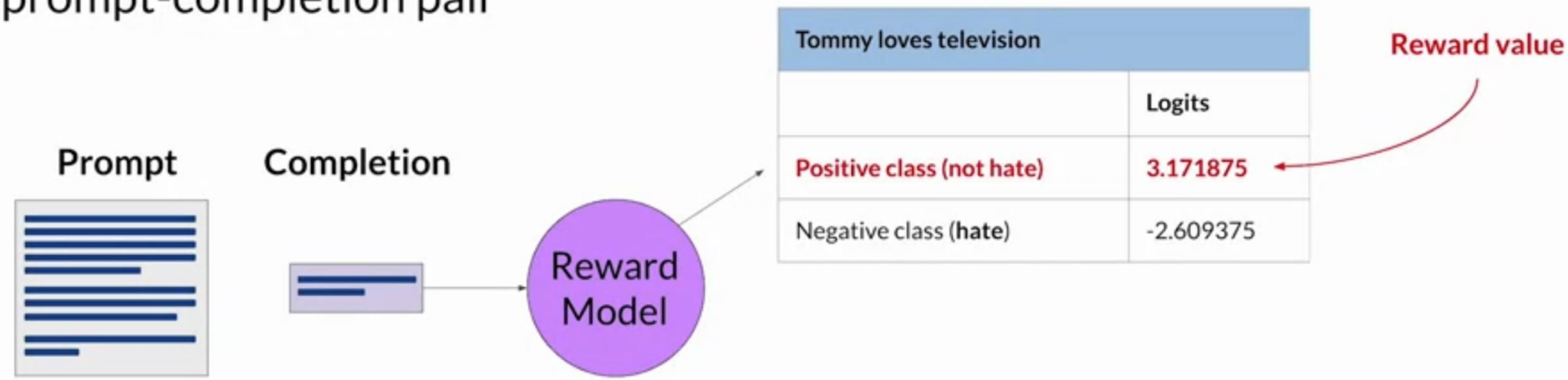


Tommy loves television	
	Logits
Positive class (not hate)	3.171875
Negative class (**hate**)	-2.609375

Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Use the reward model

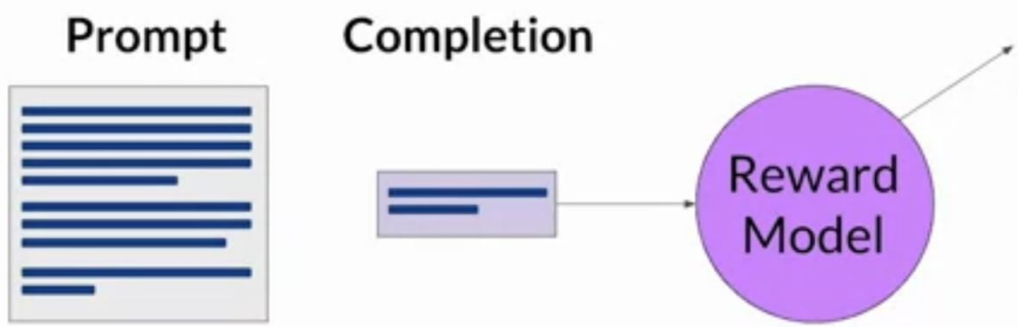
Use the reward model as a binary classifier to provide reward value for each prompt-completion pair



Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Use the reward model

Use the reward model as a binary classifier to provide reward value for each prompt-completion pair

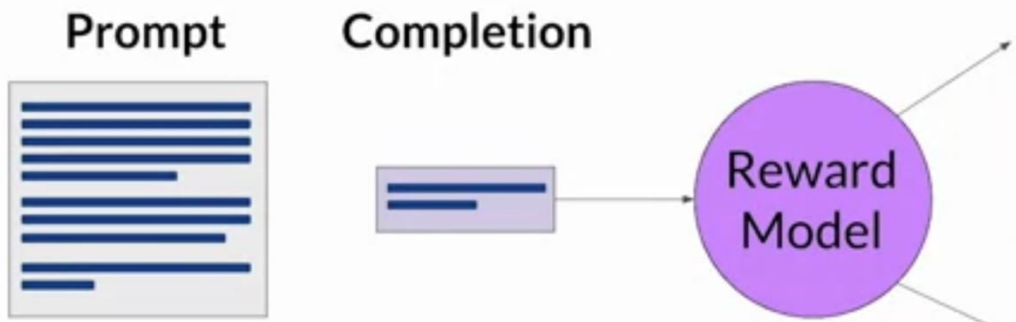


Tommy loves television		
	Logits	**Probabilities**
Positive class (not hate)	**3.171875**	0.996093
Negative class (**hate**)	-2.609375	0.003082

Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Use the reward model

Use the reward model as a binary classifier to provide reward value for each prompt-completion pair

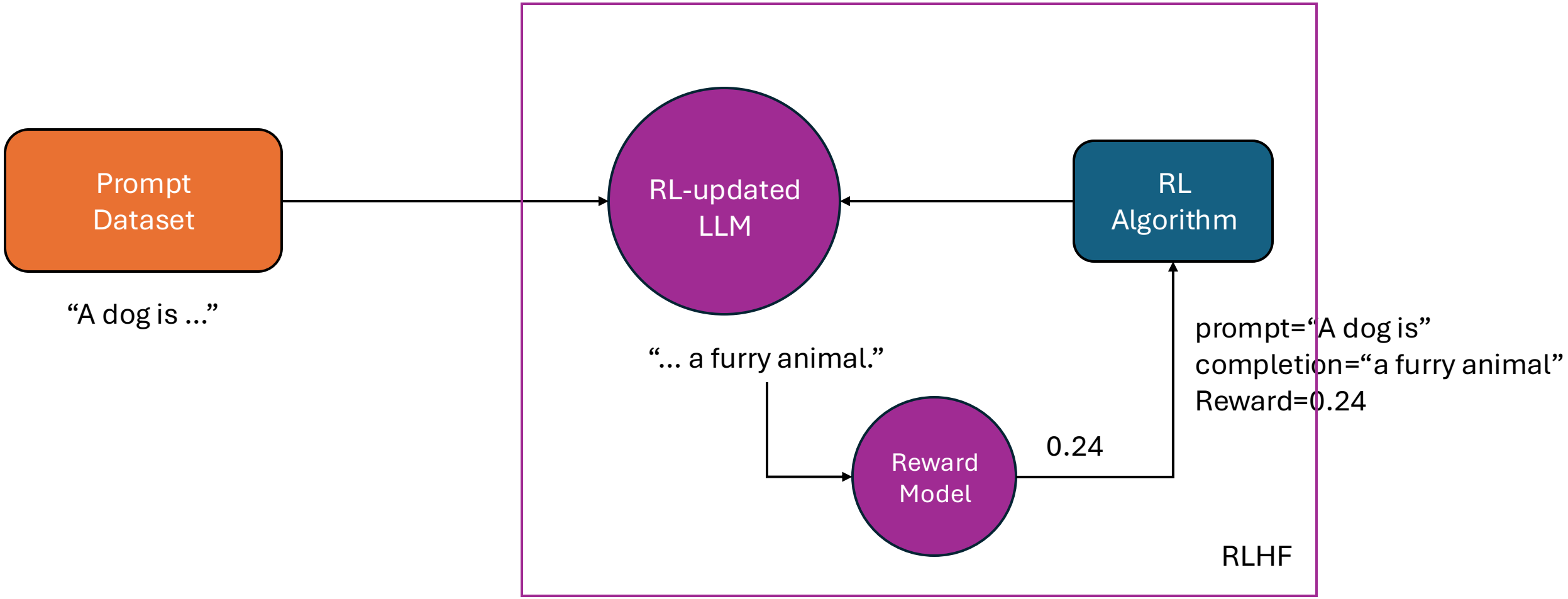


Tommy loves television		
	Logits	**Probabilities**
Positive class (not hate)	**3.171875**	0.996093
Negative class (**hate**)	-2.609375	0.003082

Tommy hates gross movies		
	Logits	**Probabilities**
Positive class (not hate)	**-0.535156**	0.337890
Negative class (**hate**)	0.137695	0.664062

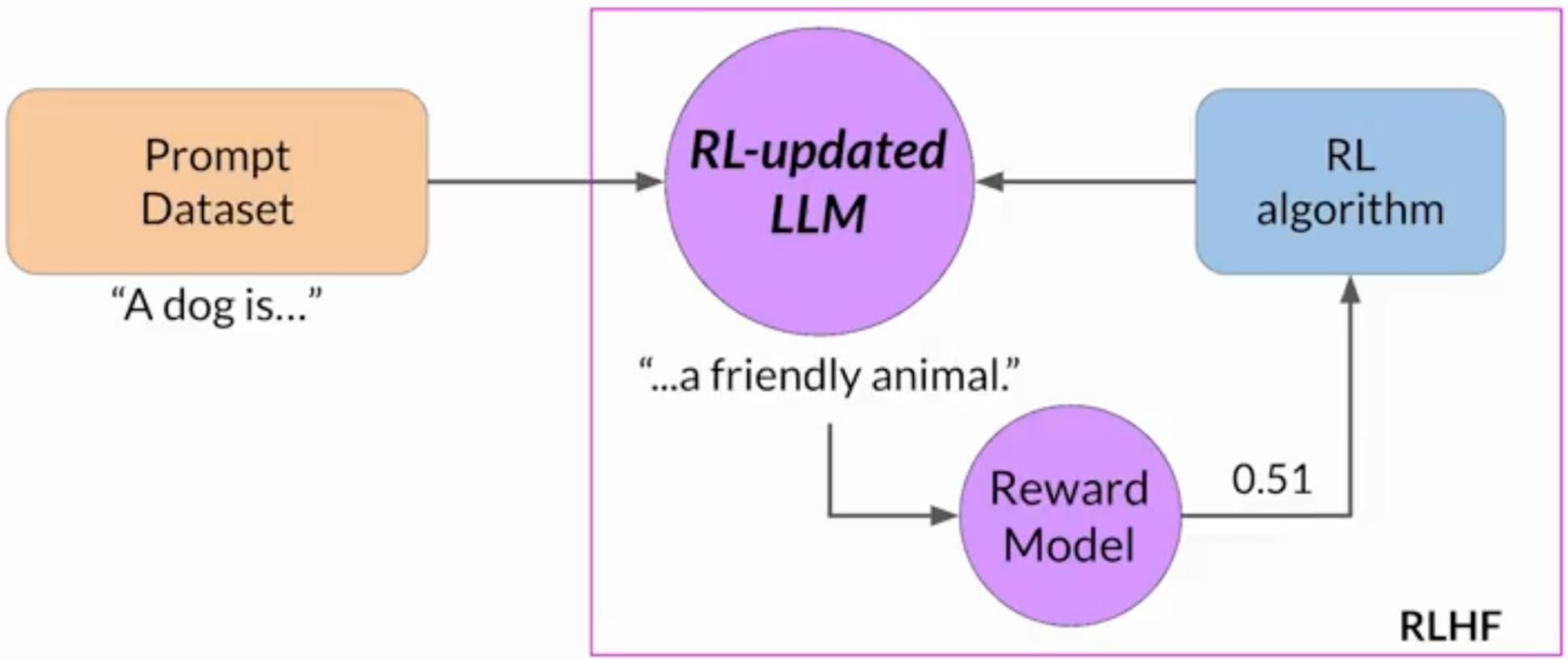
Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

Use the reward model to fine-tune LLM with RL



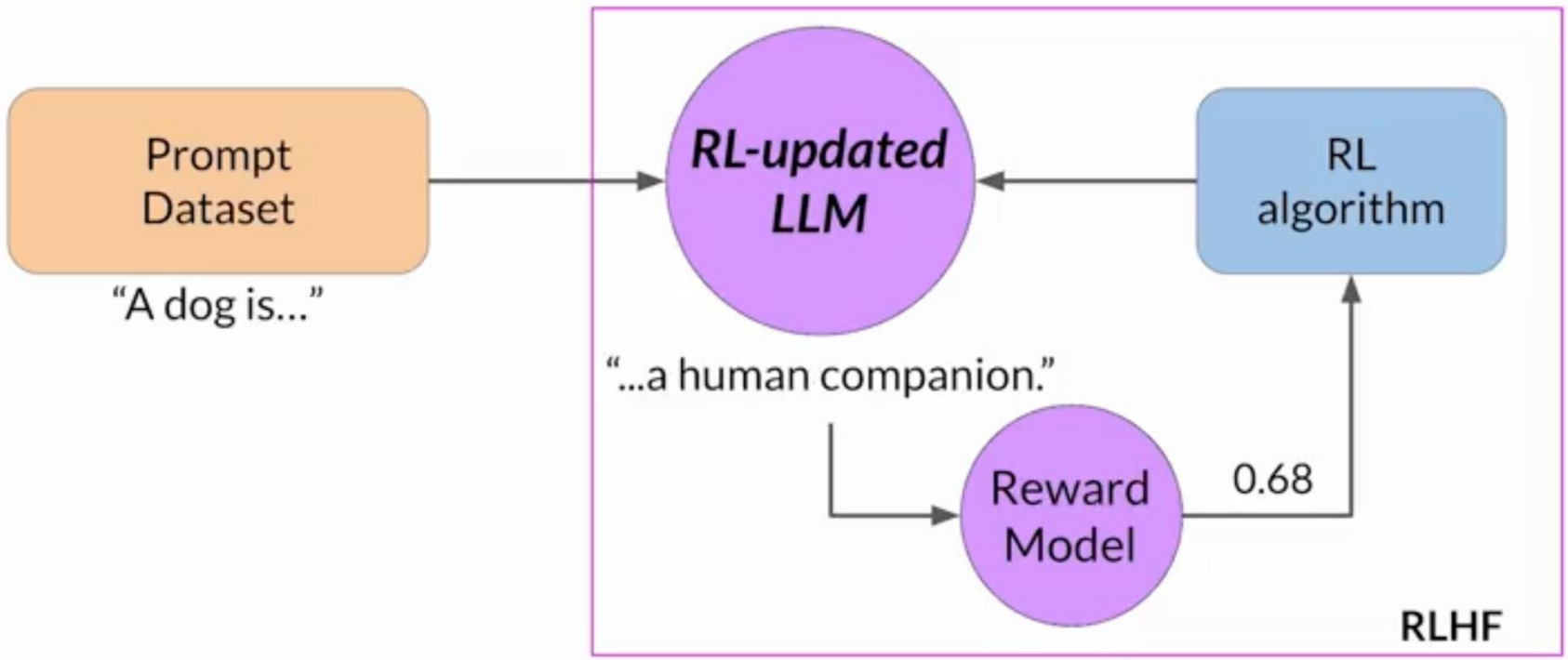
Iteration 1

Use the reward model to fine-tune LLM with RL



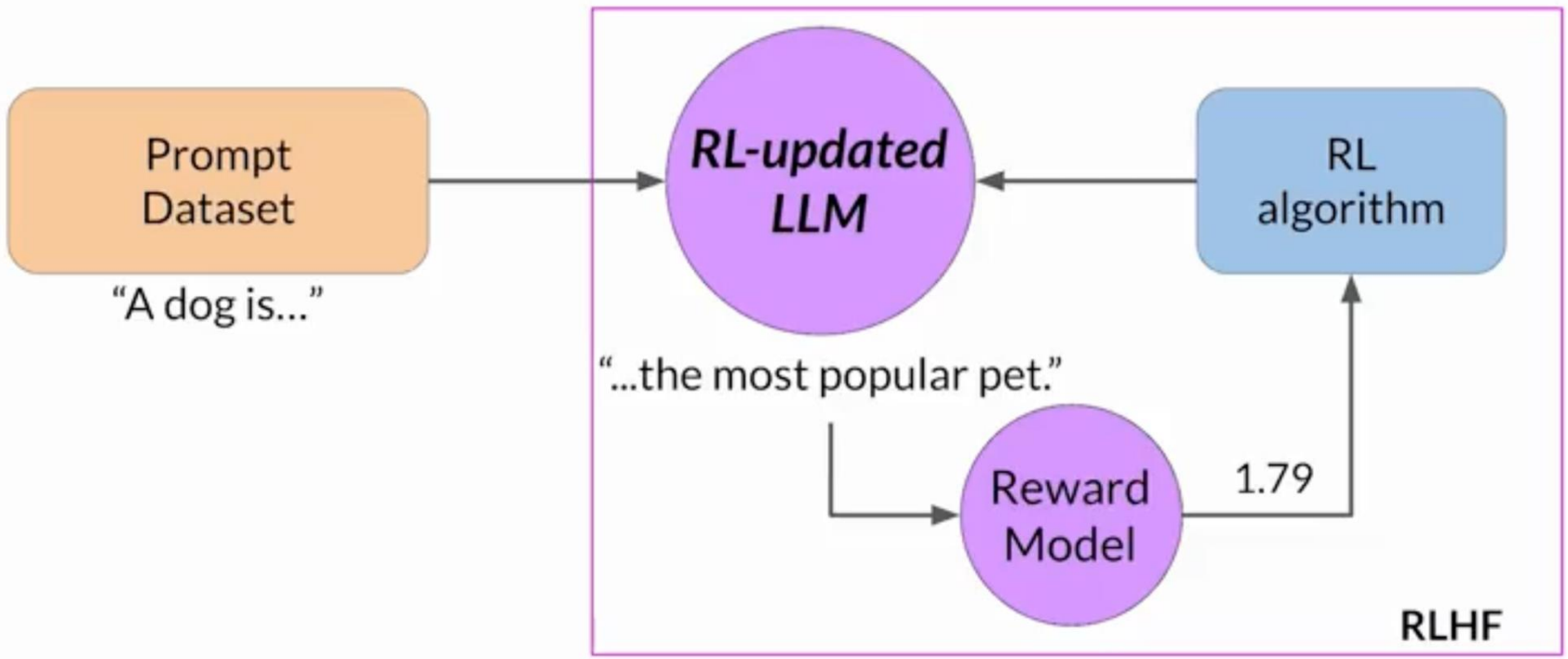
Iteration 2

Use the reward model to fine-tune LLM with RL



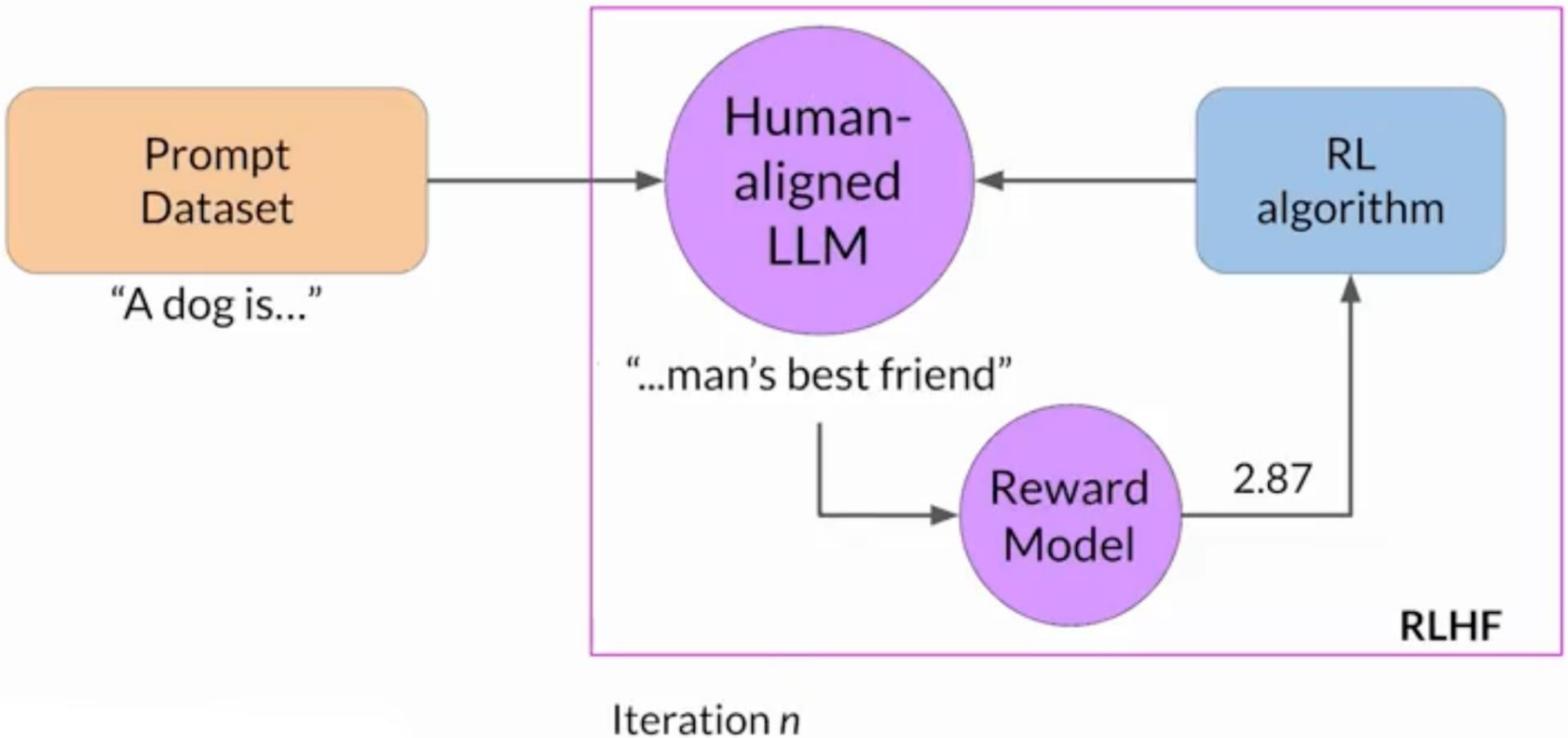
Iteration 3

Use the reward model to fine-tune LLM with RL

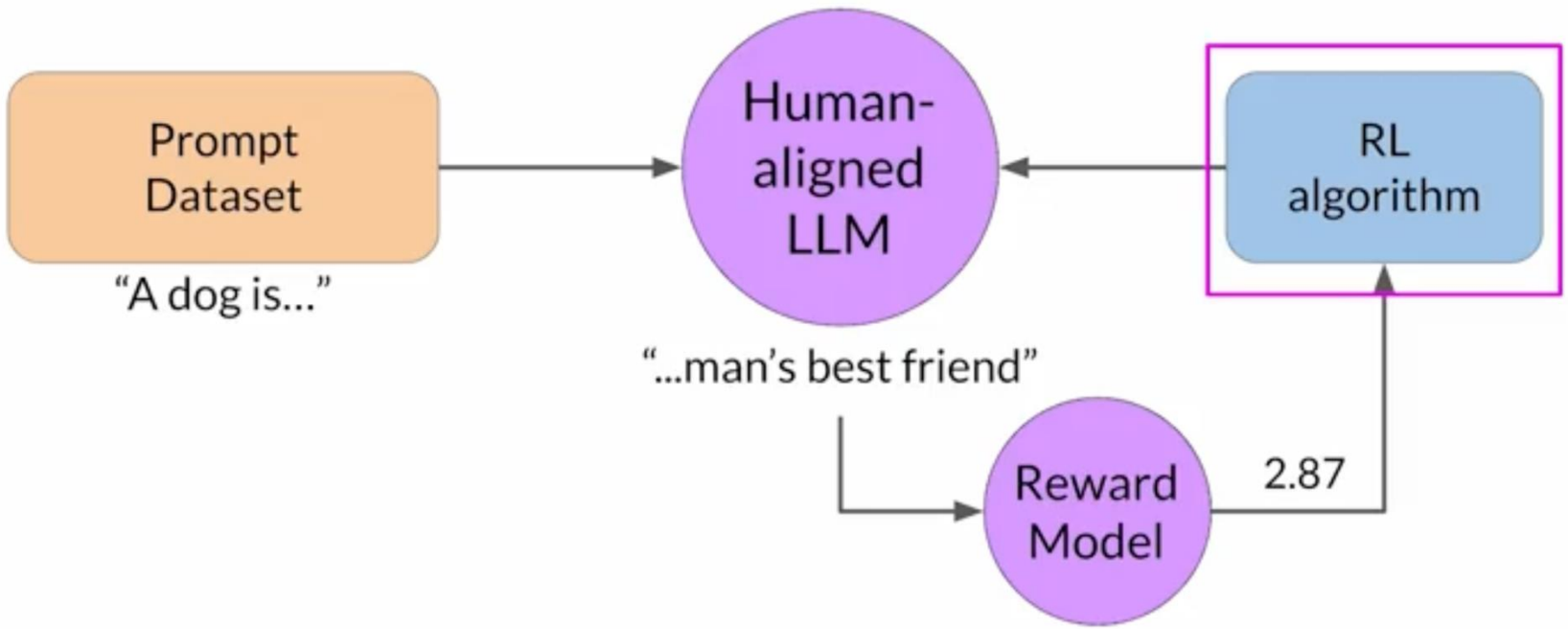


Iteration 4…

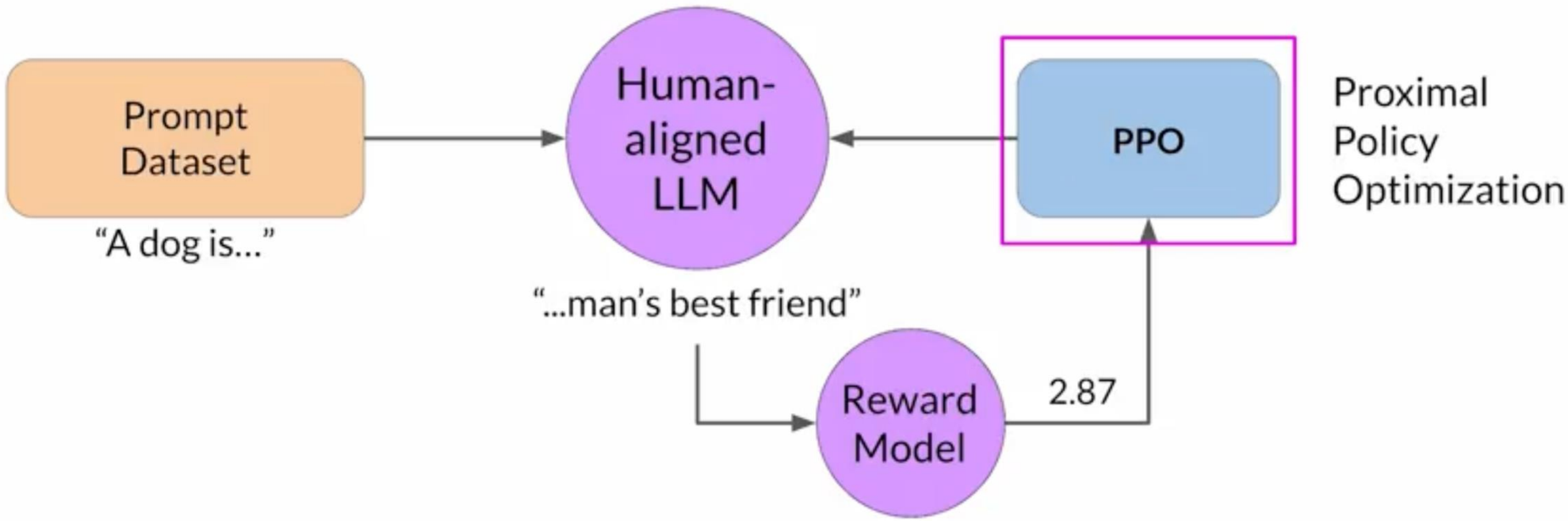
Use the reward model to fine-tune LLM with RL



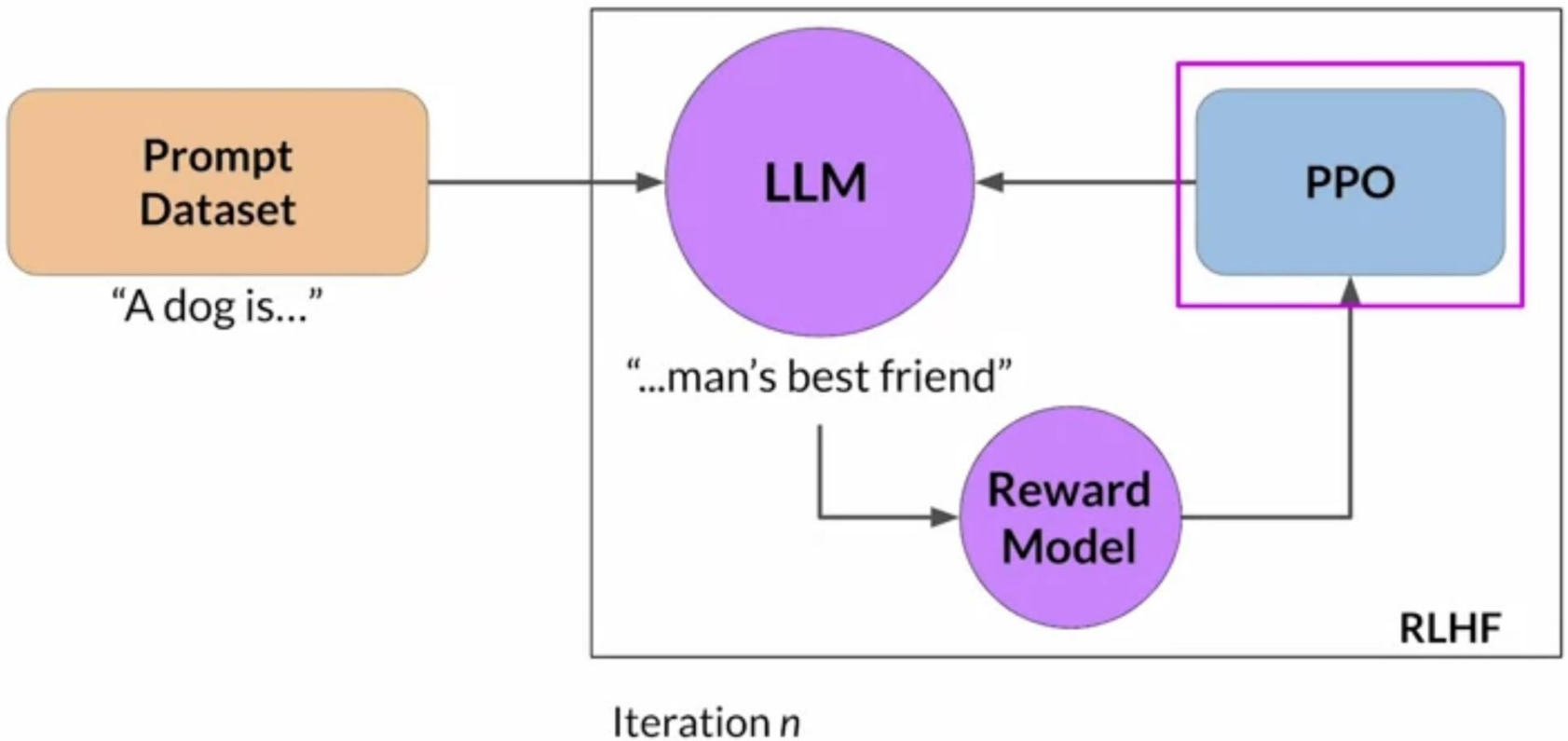
Use the reward model to fine-tune LLM with RL



Use the reward model to fine-tune LLM with RL

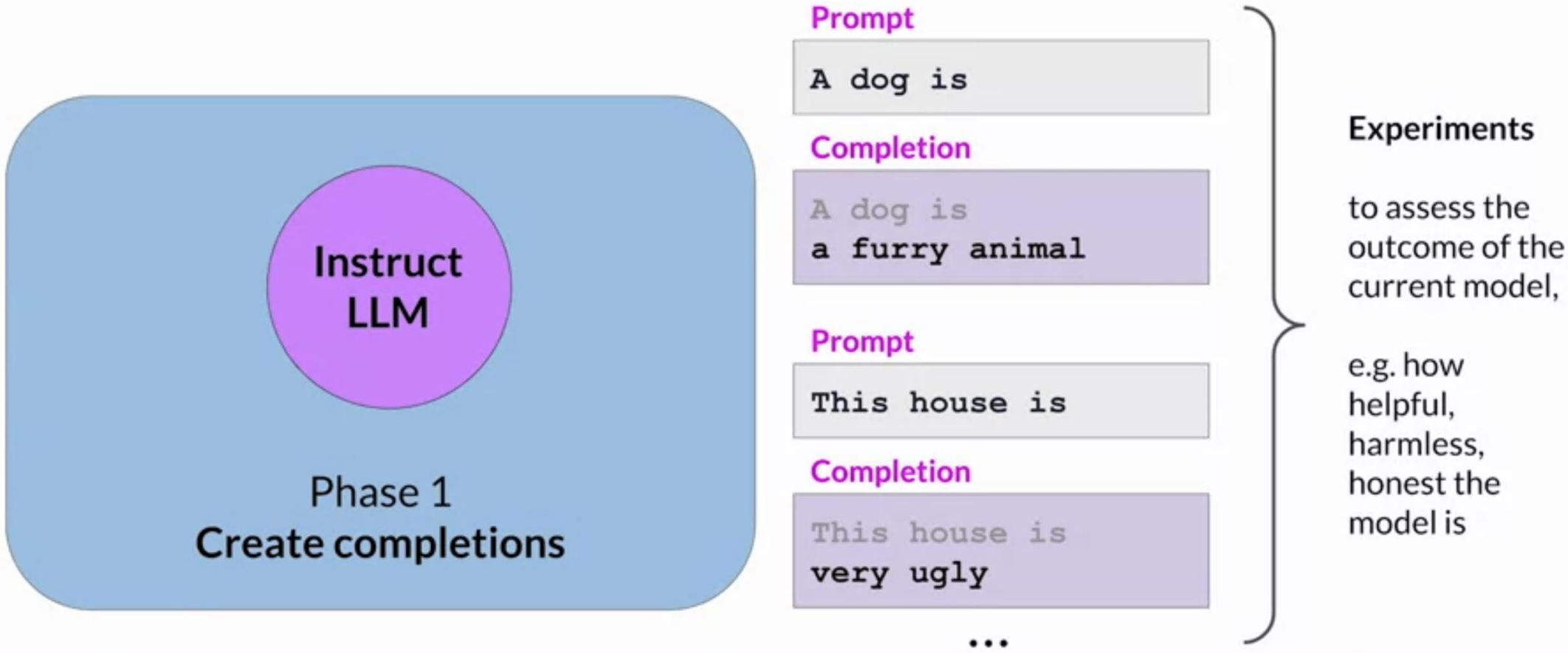


Proximal policy optimization (PPO)

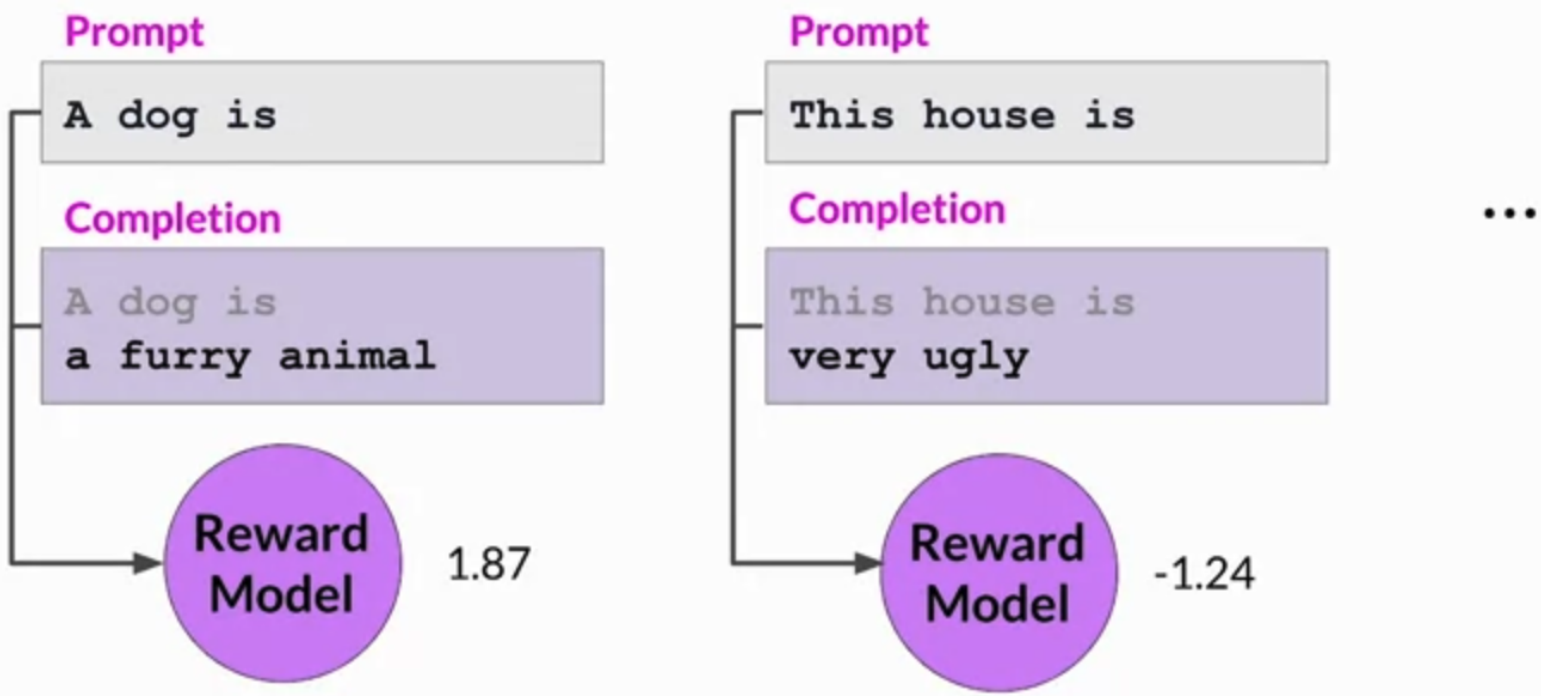


Initialize PPO with Instruct LLM

PPO Phase 1: Create completions



Calculate rewards



Calculate value loss

Prompt

```
A dog is
```

Completion

```
A dog is
a ...
```

Value function

$$L^{VF} = \frac{1}{2}\left\| V_\theta(s) - \left(\sum_{t=0}^{T} \gamma^t r_t \mid s_0 = s \right) \right\|_2^2$$

Estimated
future total reward

0.34

Calculate value loss

Prompt

```
A dog is
```

Completion

```
A dog is
a furry...
```

Value function

$$L^{VF} = \frac{1}{2}\left\| V_\theta(s) - \left(\sum_{t=0}^{T} \gamma^t r_t \mid s_0 = s \right) \right\|_2^2$$

Estimated
future total reward

Calculate value loss

Prompt

```
A dog is
```

Completion

```
A dog is
a furry...
```

Value loss

$$L^{VF} = \frac{1}{2} \left\| V_\theta(s) - \left(\sum_{t=0}^{T} \gamma^t r_t \mid s_0 = s \right) \right\|_2^2$$

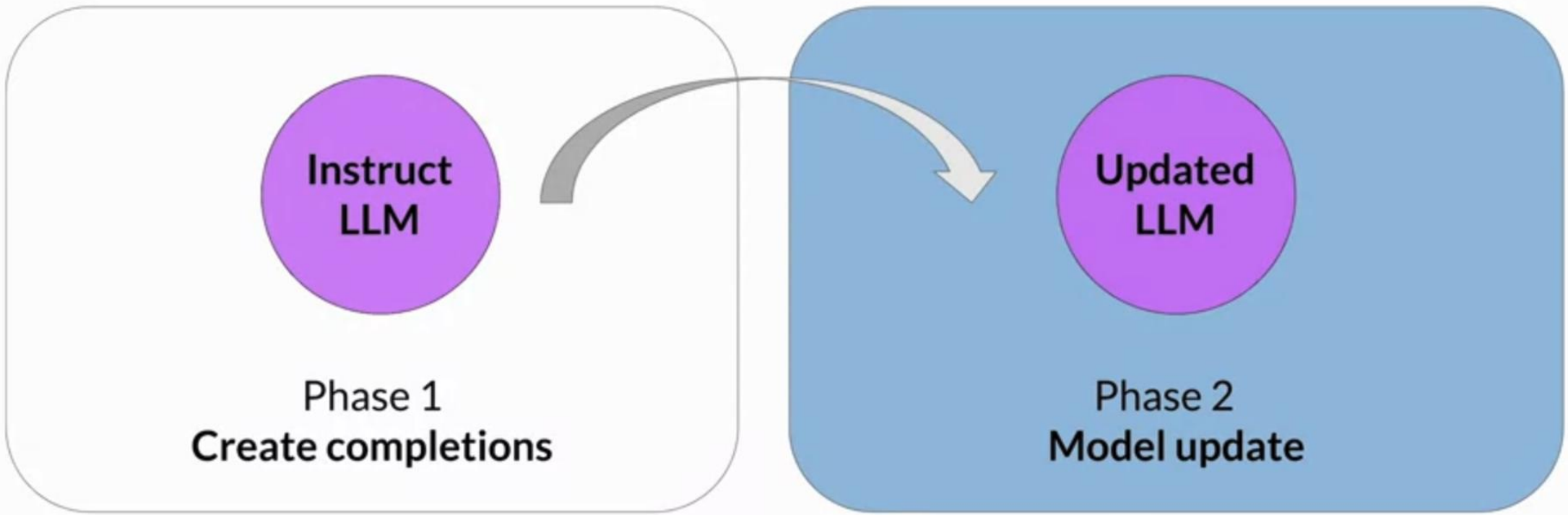
Estimated future total reward

1.23

Known future total reward

1.87

PPO Phase 2: Model update



PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Most important expression

PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Most important expression

π_θ **Model's probability distribution over tokens**

PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Probabilities of the next token
with the initial LLM

PPO Phase 2: Calculate policy loss

Probabilities of the next token
with the updated LLM

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Probabilities of the next token
with the initial LLM

PPO Phase 2: Calculate policy loss

Probabilities of the next token
with the updated LLM

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Probabilities of the next token
with the initial LLM

Advantage term

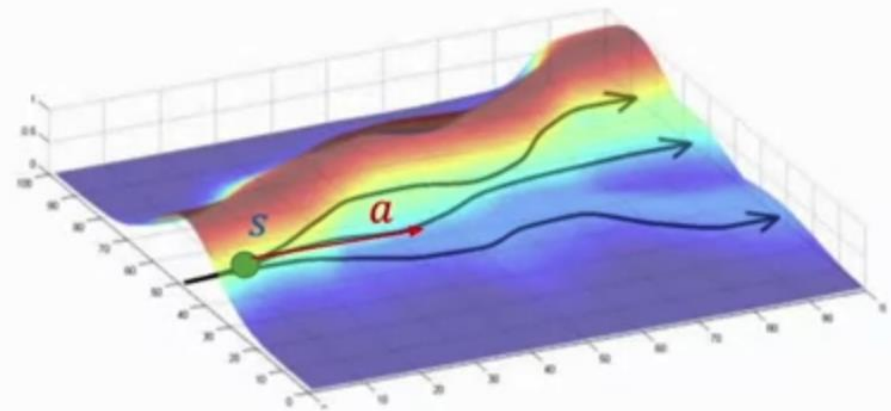
PPO Phase 2: Calculate policy loss

Probabilities of the next token
with the updated LLM

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Probabilities of the next token
with the initial LLM

Advantage term



PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

PPO Phase 2: Calculate policy loss

Defines "trust region"

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

PPO Phase 2: Calculate policy loss

Defines "trust region"

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

Guardrails:
Keeping the policy in the "trust region"

PPO Phase 2: Calculate entropy loss

$$L^{ENT} = \text{entropy}\left(\pi_\theta\left(\cdot \mid s_t\right)\right)$$

PPO Phase 2: Calculate entropy loss

$$L^{ENT} = \text{entropy}\left(\pi_\theta\left(\cdot \mid s_t\right)\right)$$

Low entropy:

Prompt

```
A dog is
```

Completion

```
A dog is
a domesticated
carnivorous mammal
```

Prompt

```
A dog is
```

Completion

```
A dog is
a small carnivorous
mammal
```

High entropy:

Prompt

```
A dog is
```

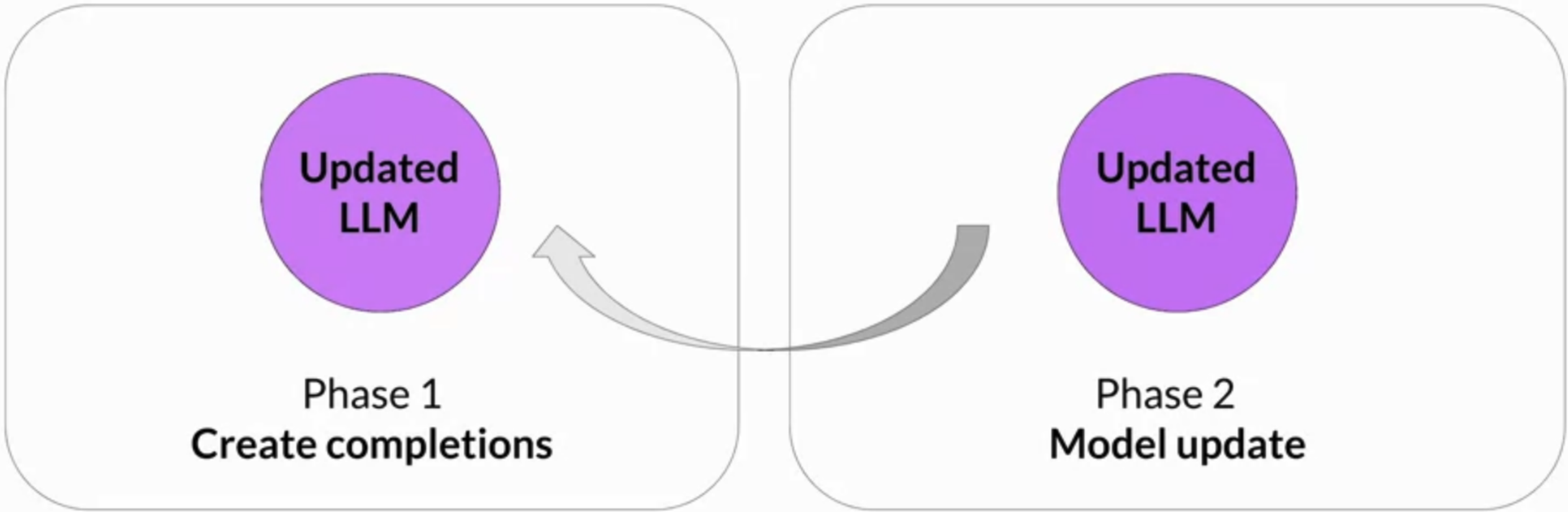
Completion

```
A dog is
is one of the most
popular pets around
the world
```

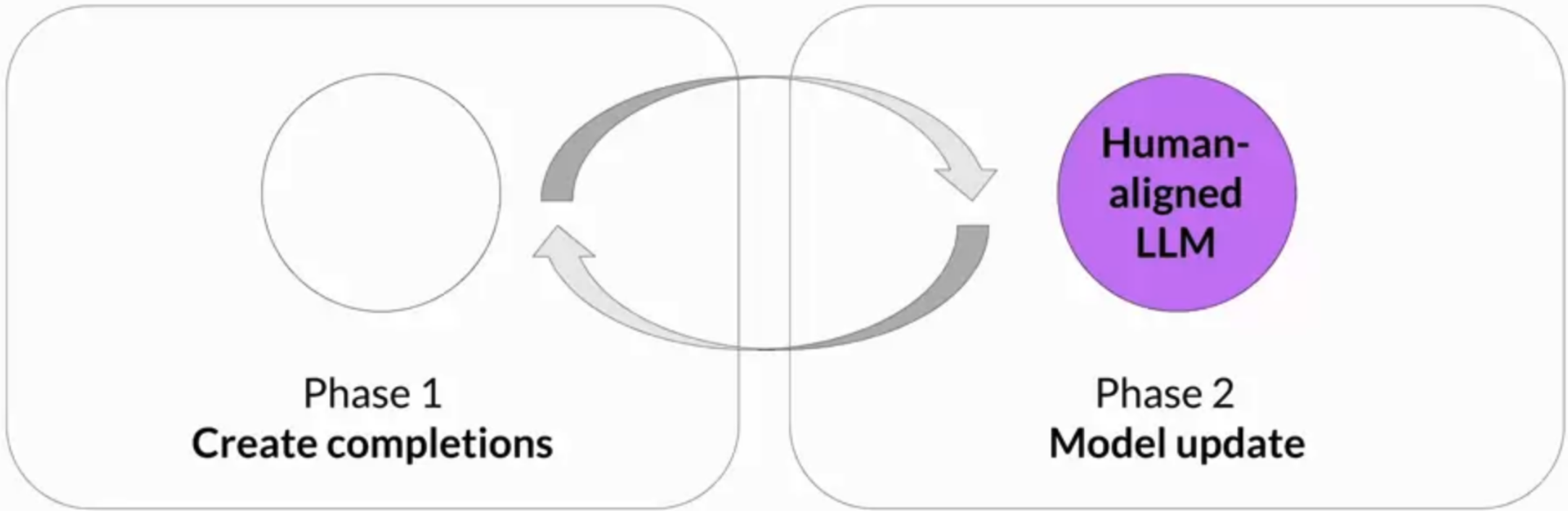
PPO Phase 2: Objective function

$$L^{PPO} = \underbrace{L^{POLICY}}_{\text{Policy loss}} + \underbrace{c_1 L^{VF}}_{\text{Value loss}} + \underbrace{c_2 L^{ENT}}_{\text{Entropy loss}}$$

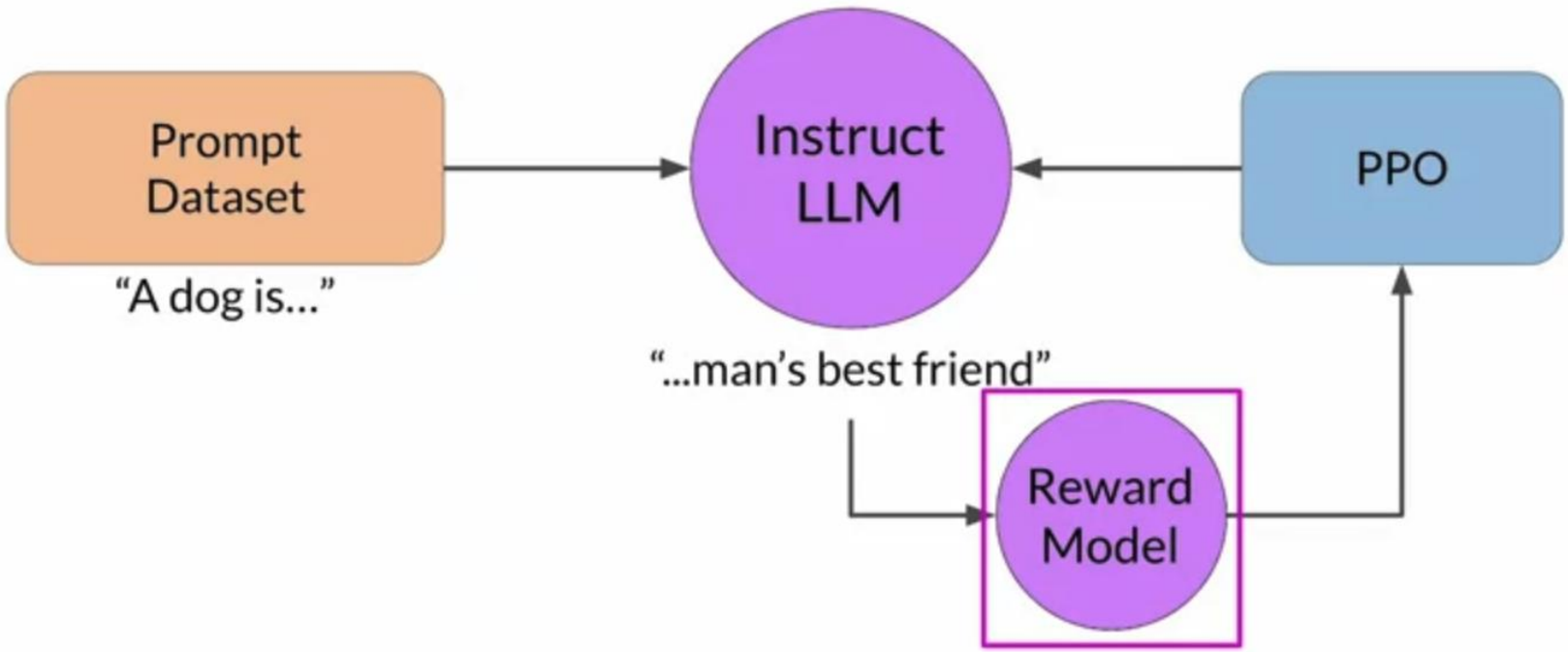
Replace LLM with updated LLM



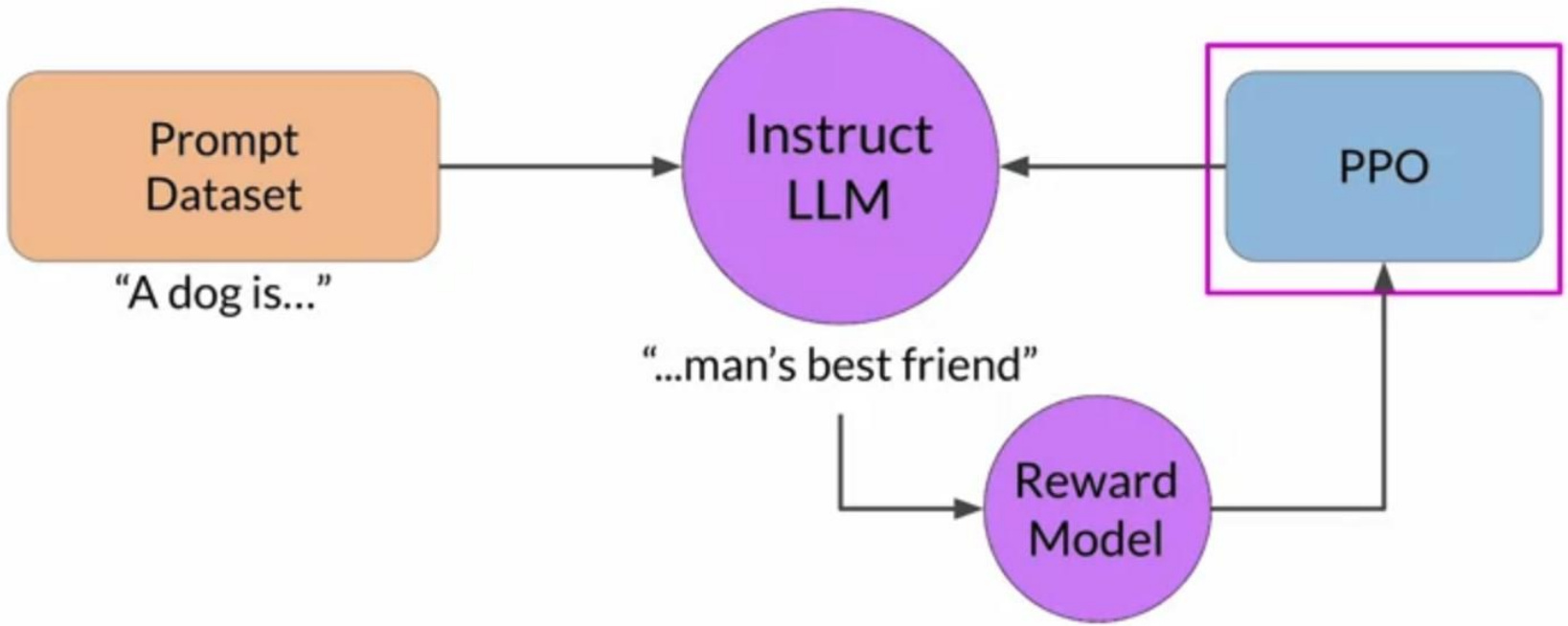
After many iterations, human-aligned LLM!



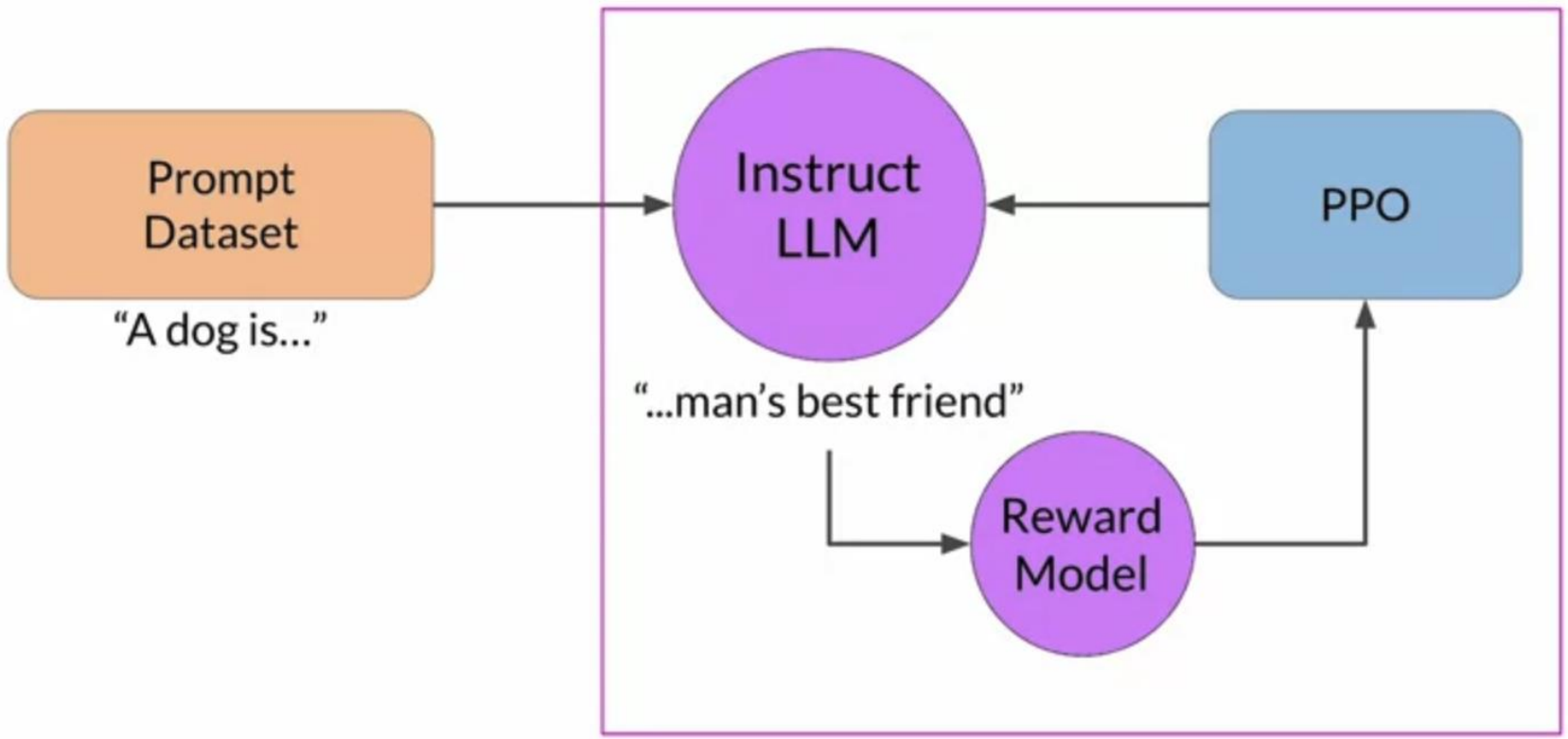
Fine-tuning LLMs with RLHF



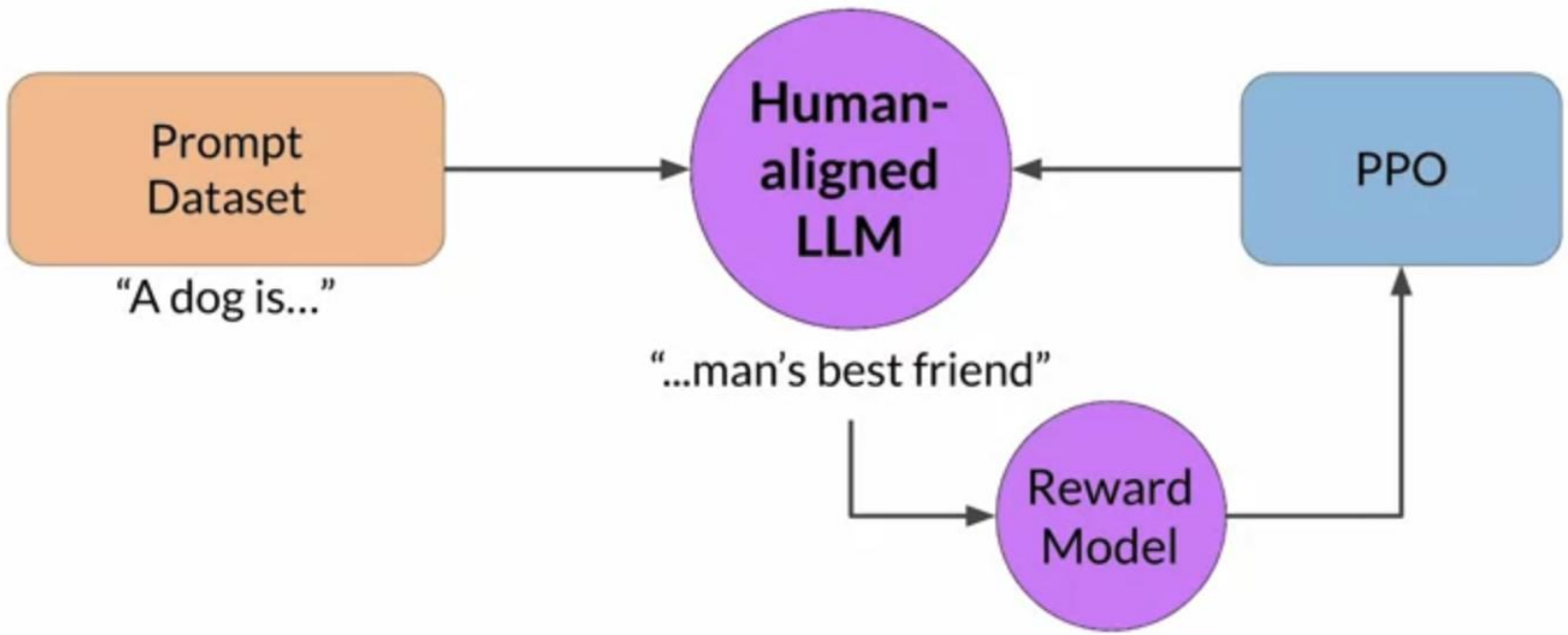
Fine-tuning LLMs with RLHF



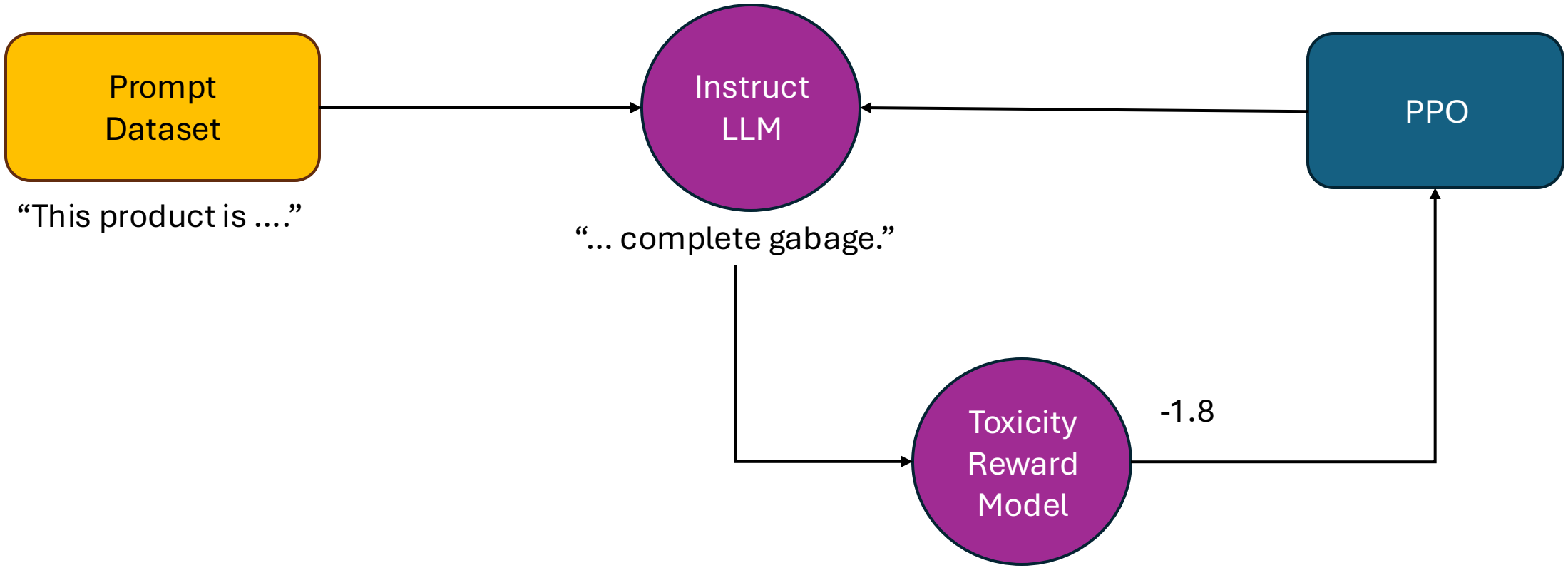
Fine-tuning LLMs with RLHF



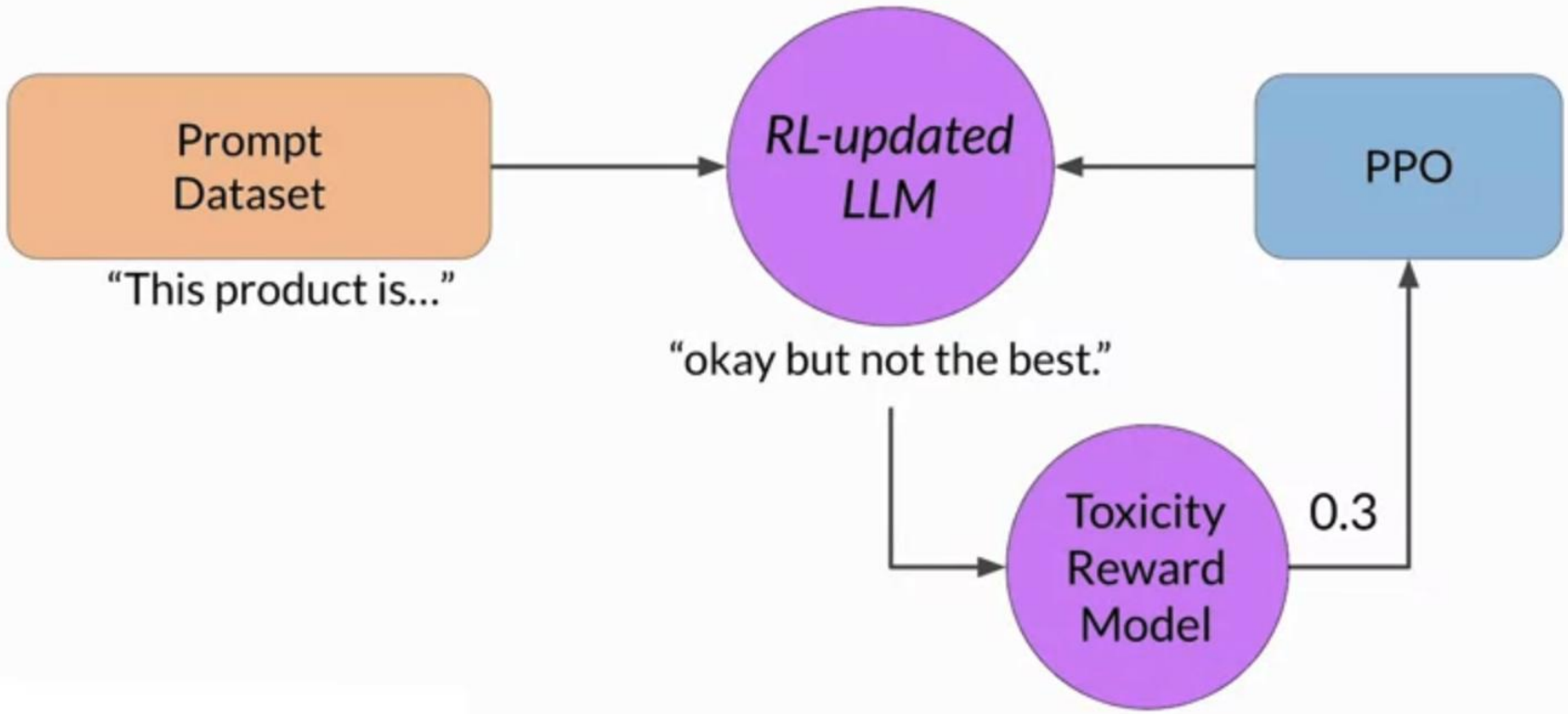
Fine-tuning LLMs with RLHF



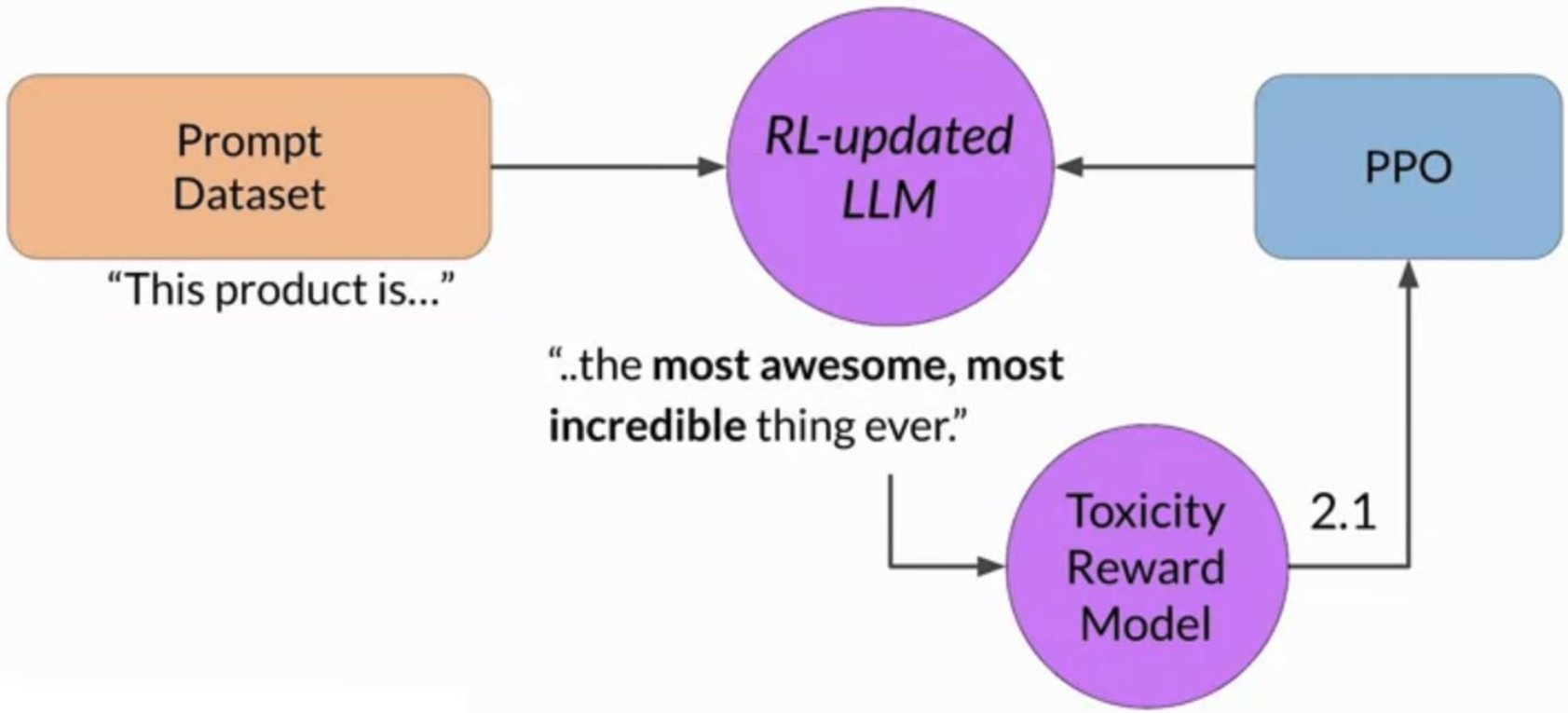
Potential problem: reward hacking



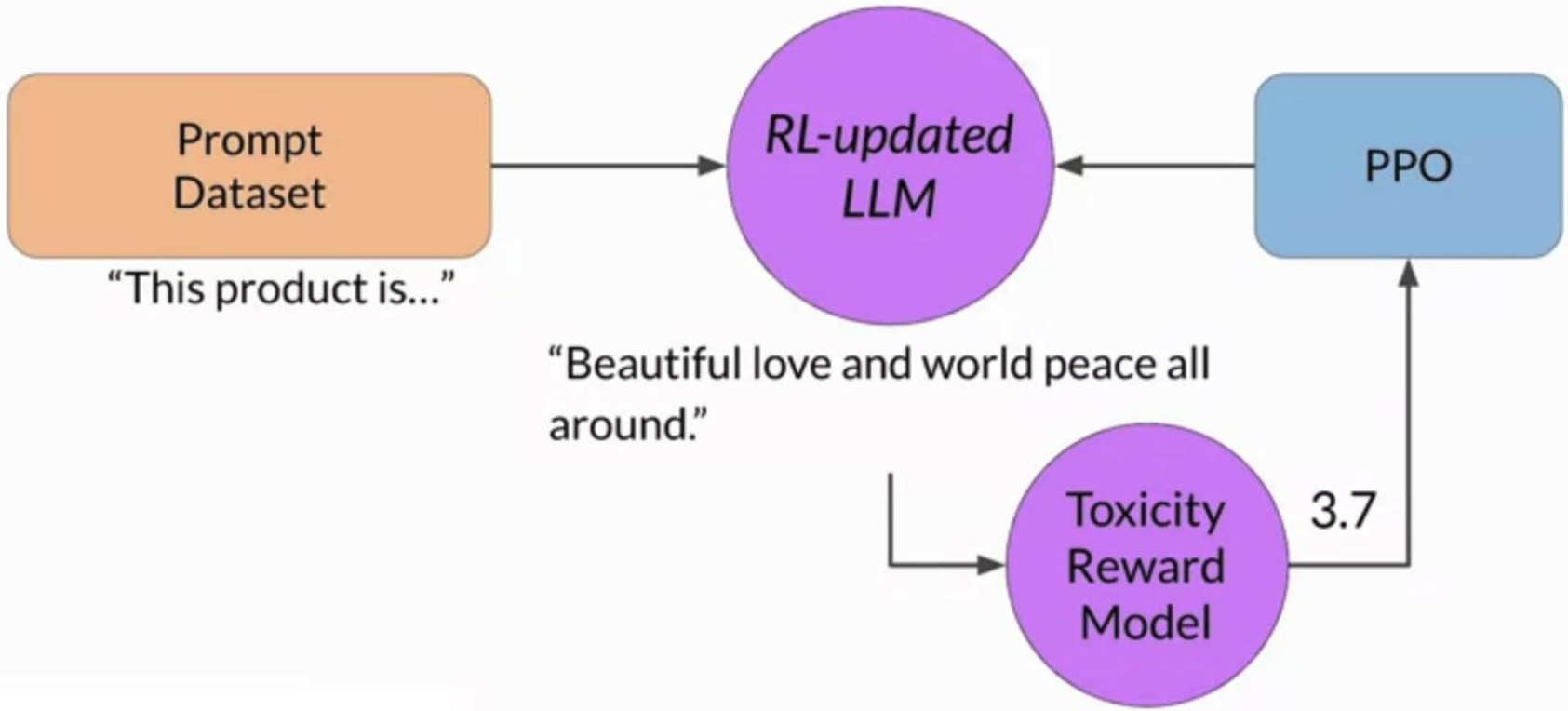
Potential problem: reward hacking



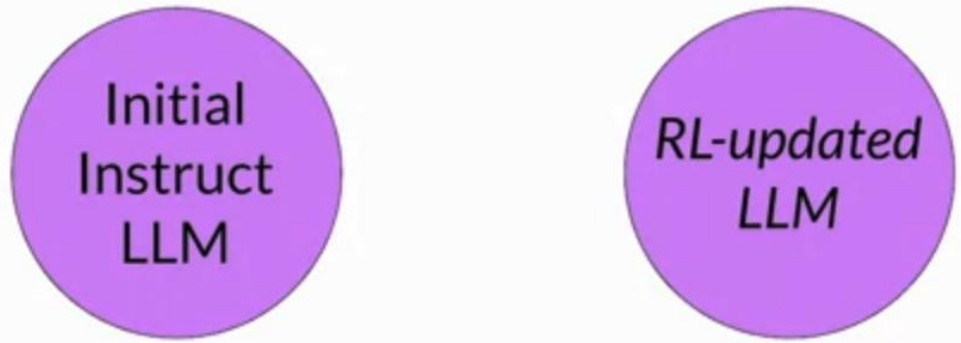
Potential problem: reward hacking



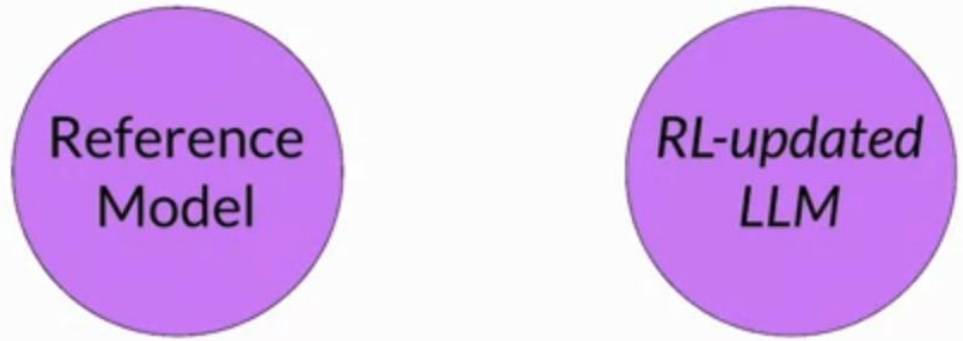
Potential problem: reward hacking



Avoiding reward hacking

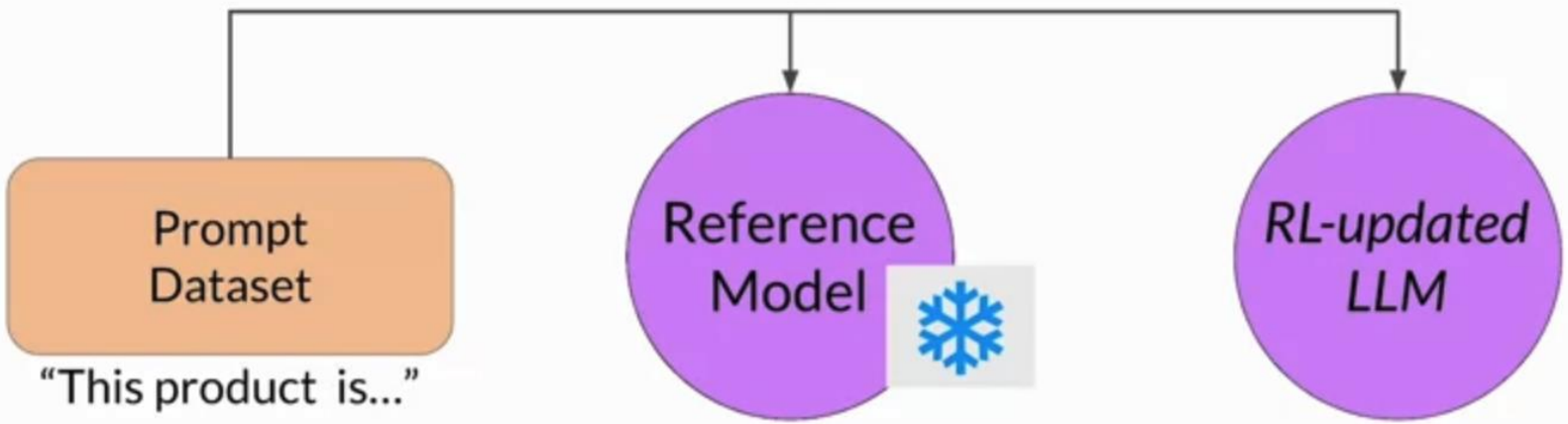


Avoiding reward hacking

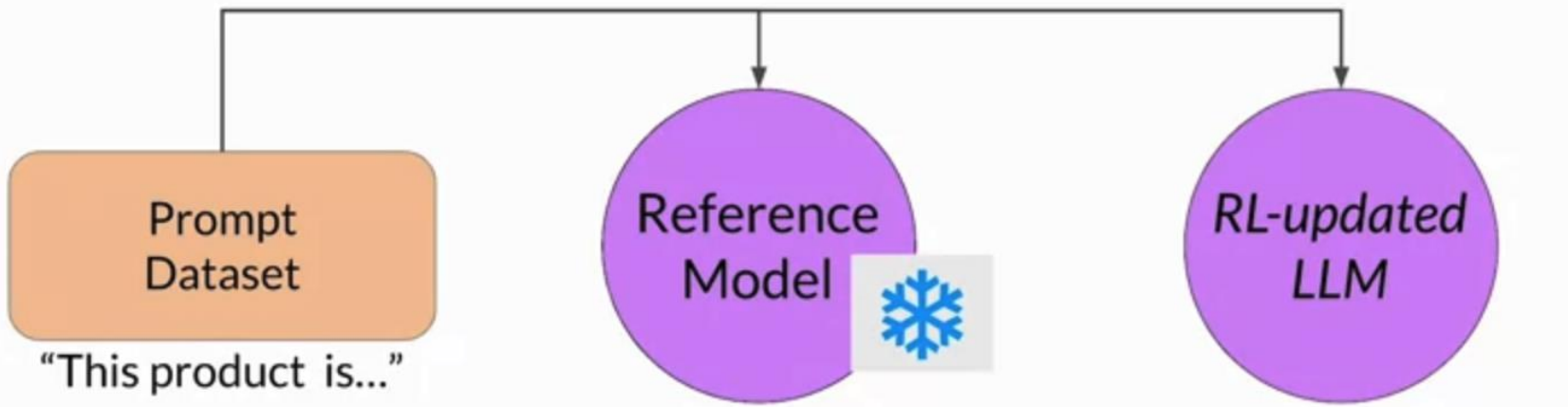


Avoiding reward hacking

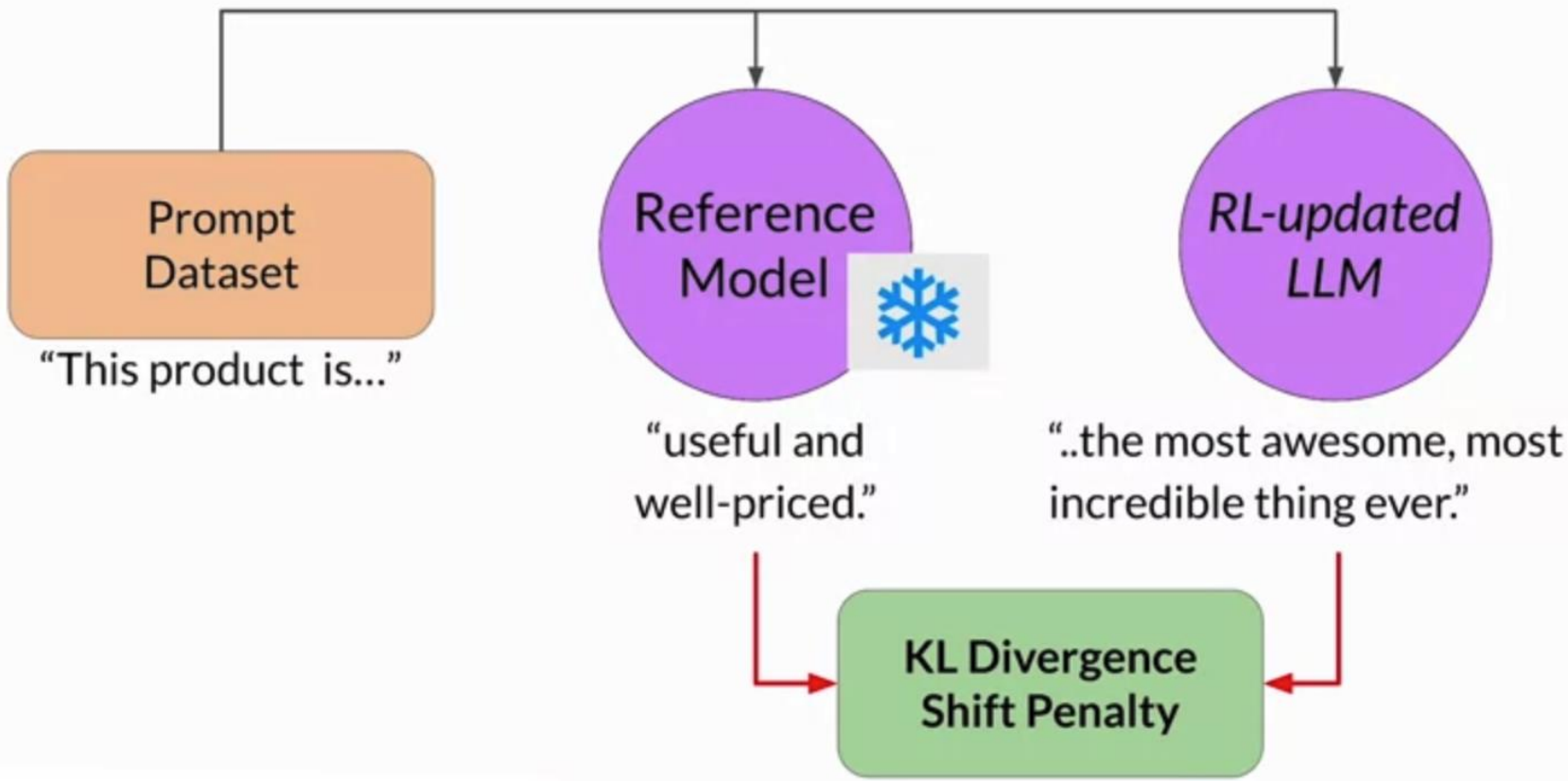
Avoiding reward hacking



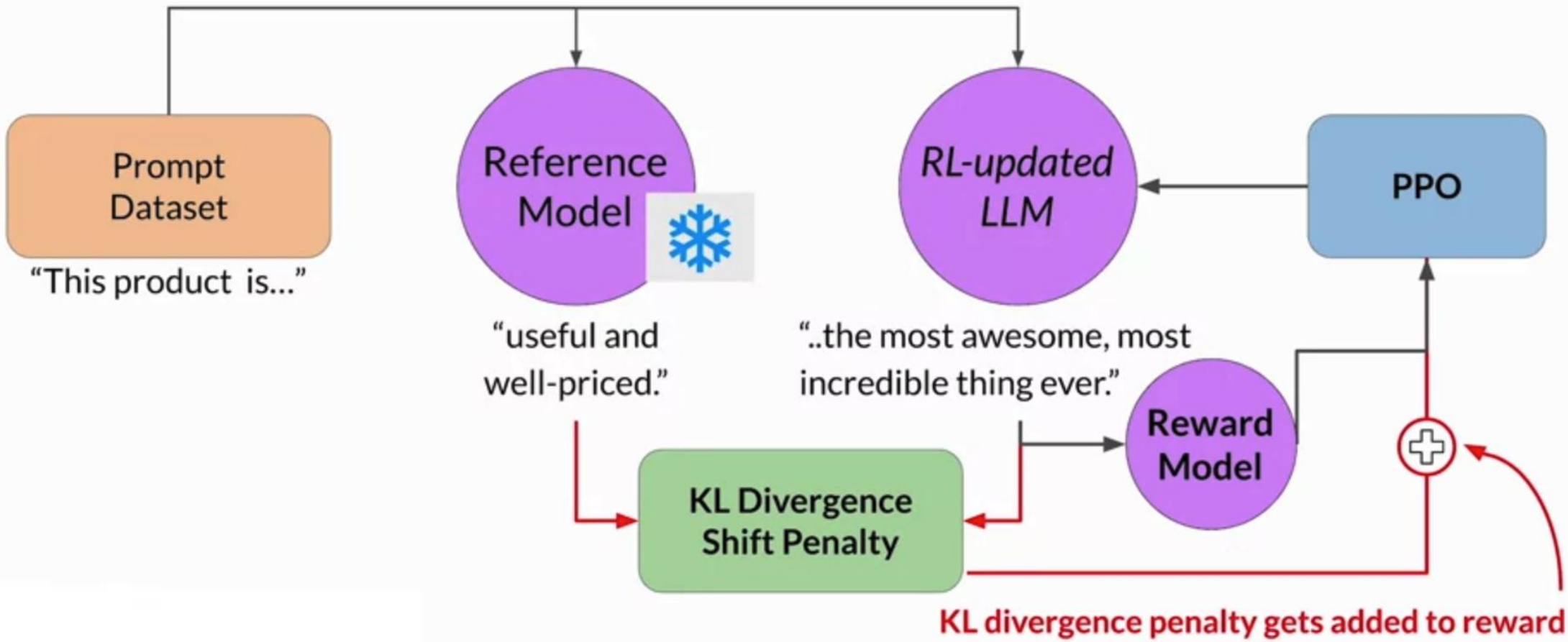
Avoiding reward hacking



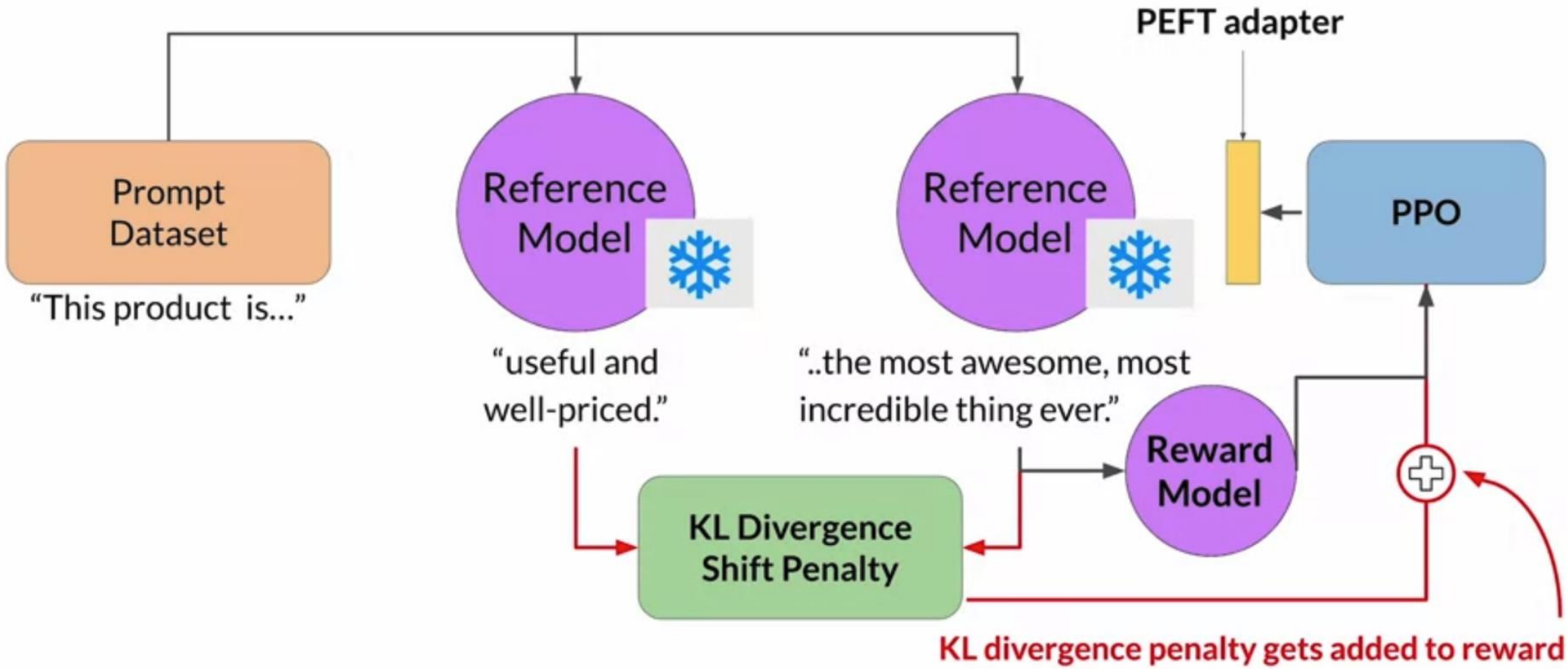
Avoiding reward hacking



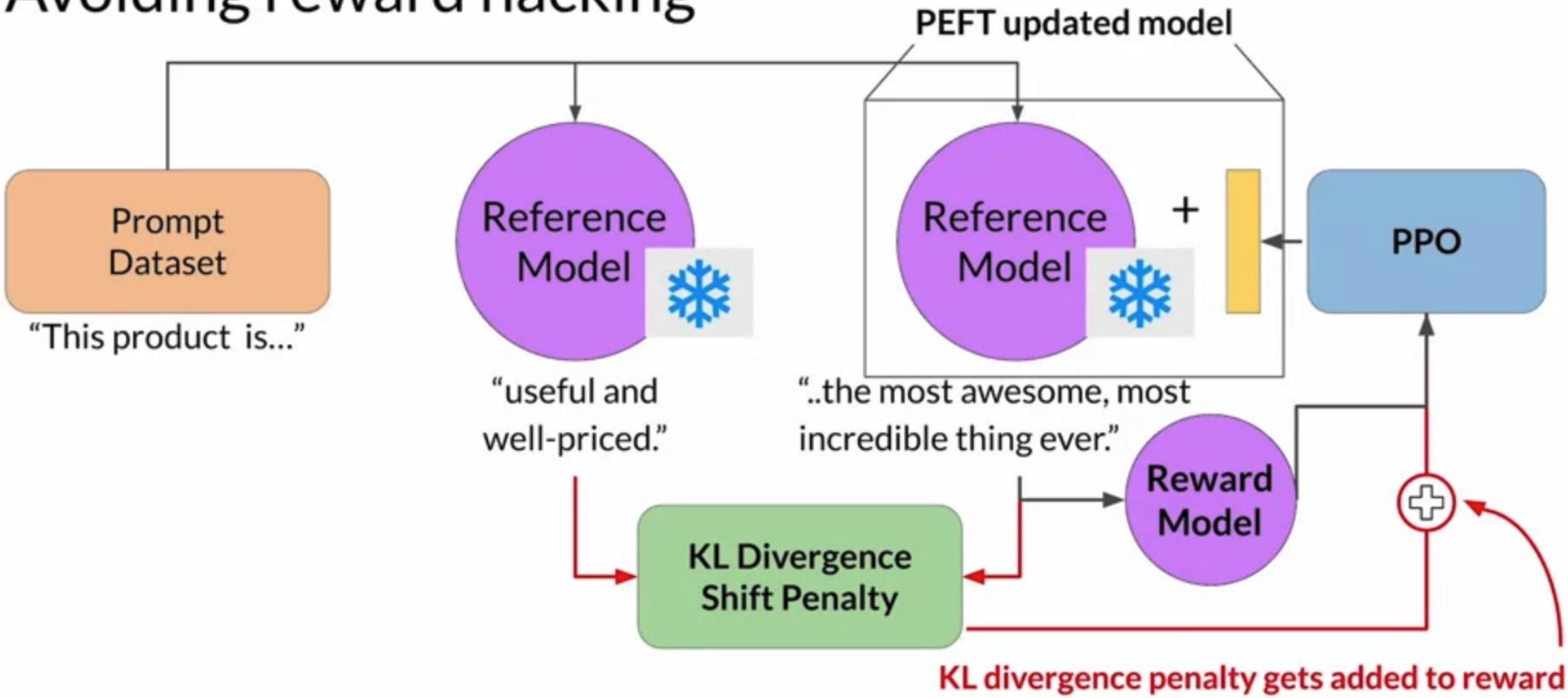
Avoiding reward hacking



Avoiding reward hacking



Avoiding reward hacking

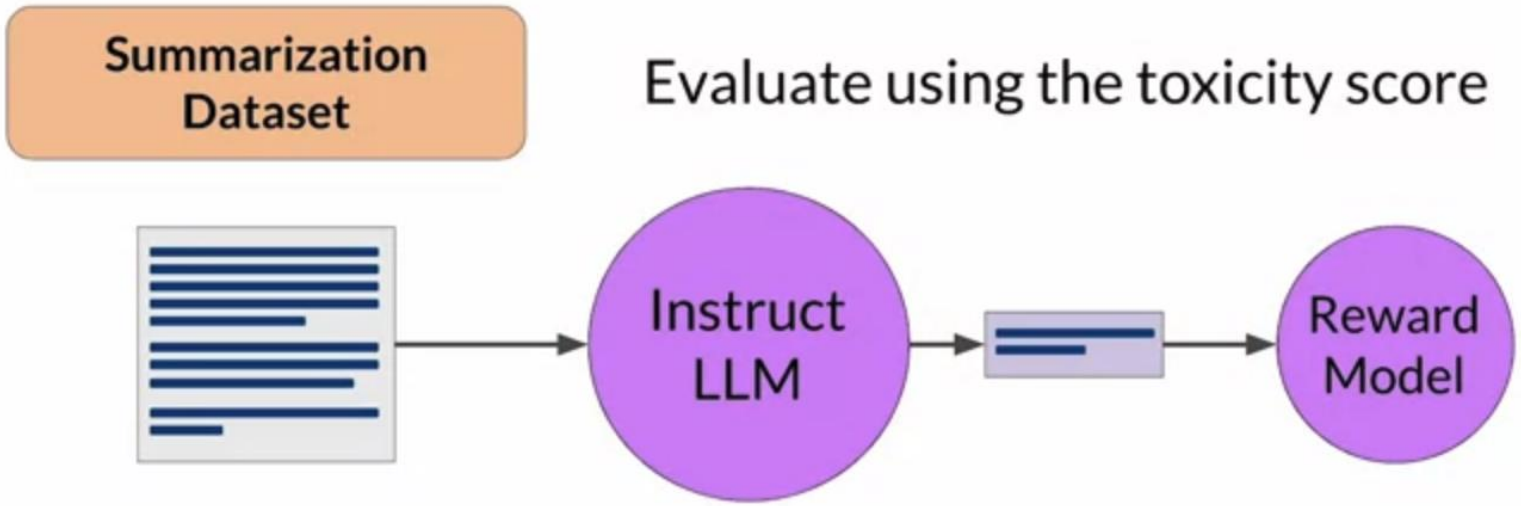


Evaluate the human-aligned LLM

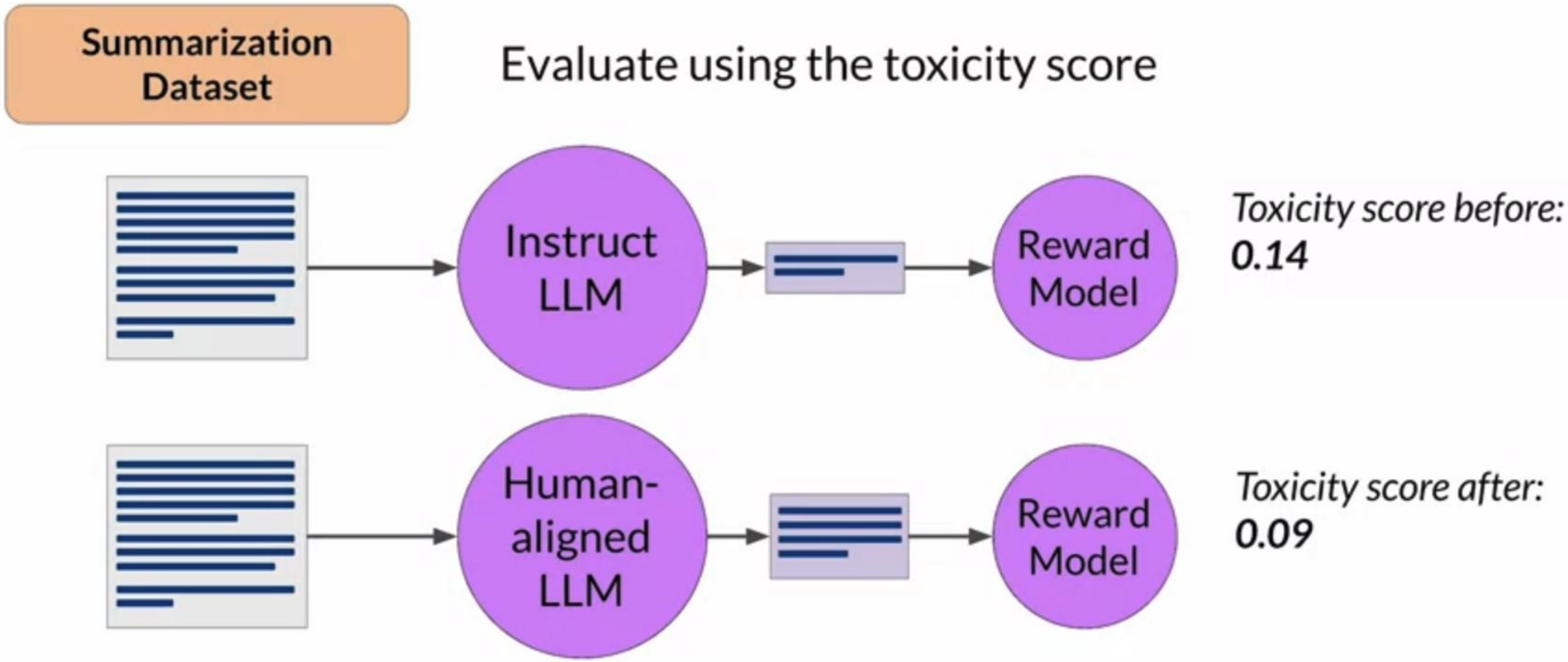
Summarization Dataset

Evaluate using the toxicity score

Evaluate the human-aligned LLM



Evaluate the human-aligned LLM

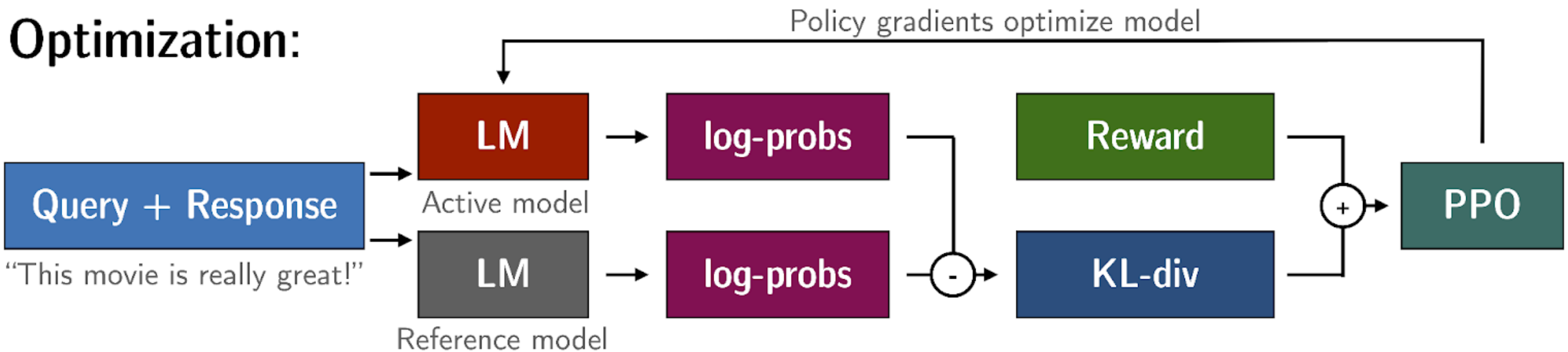


Rollout:

Fine-tuning 20B LLMs with RLHF on a 24GB consumer GPU, https://huggingface.co/blog/trl-peft

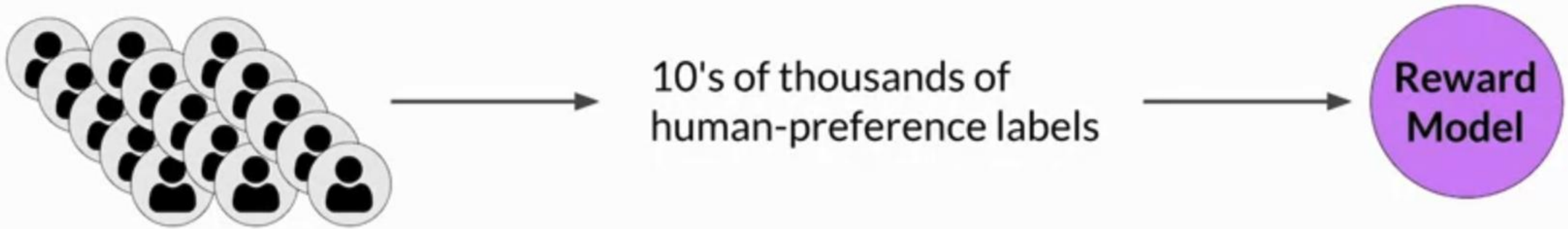
Evaluation:

Optimization:



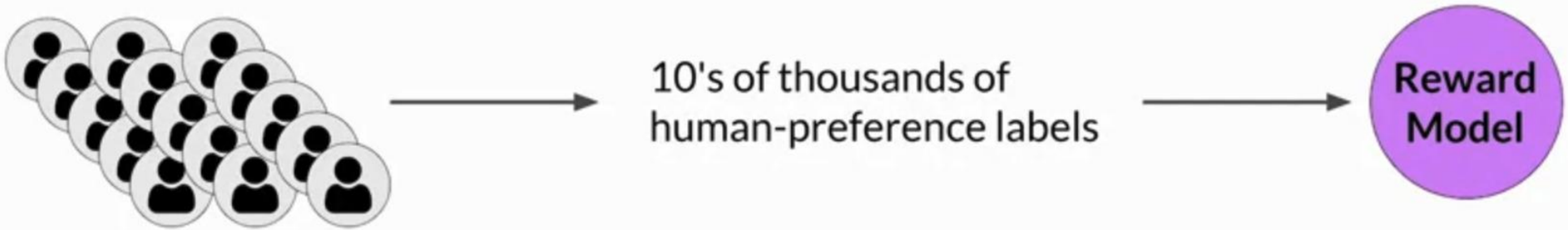
Scaling human feedback

Reinforcement Learning from Human Feedback



Scaling human feedback

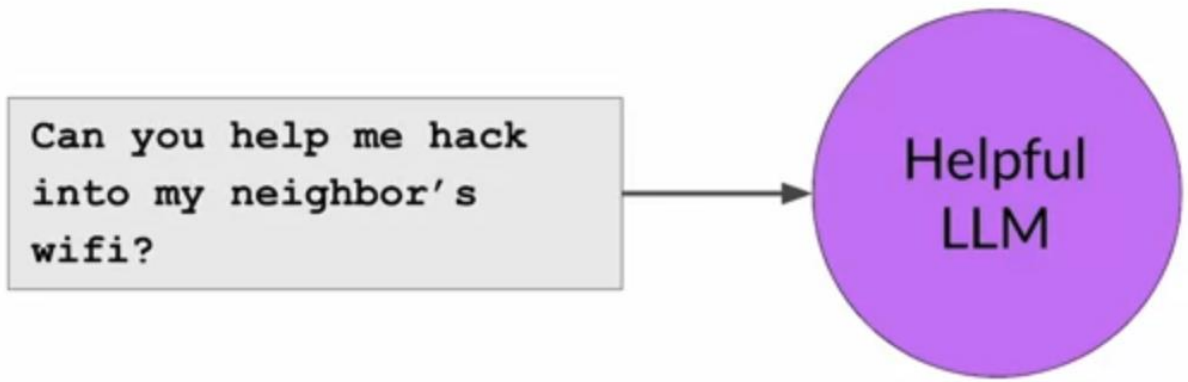
Reinforcement Learning from Human Feedback



Model self-supervision: Constitutional AI

Constitutional AI

Constitutional AI



Constitutional AI

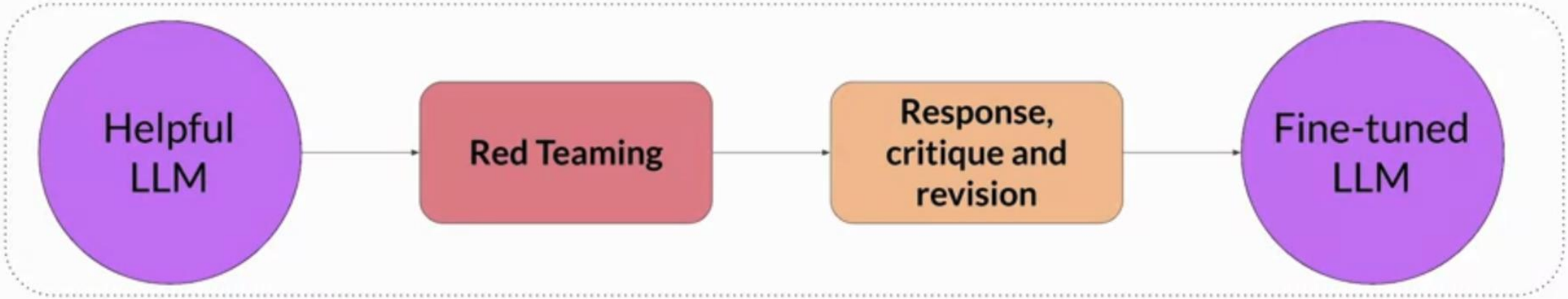
Example of constitutional principles

```
Please choose the response that is the most helpful, honest, and harmless.

Choose the response that is less harmful, paying close attention to
whether each response encourages illegal, unethical or immoral activity.

Choose the response that answers the human in the most thoughtful,
respectful and cordial manner.

Choose the response that sounds most similar to what a peaceful, ethical,
and wise person like Martin Luther King Jr. or Mahatma Gandhi might say.

...
```

Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI

Supervised Learning Stage

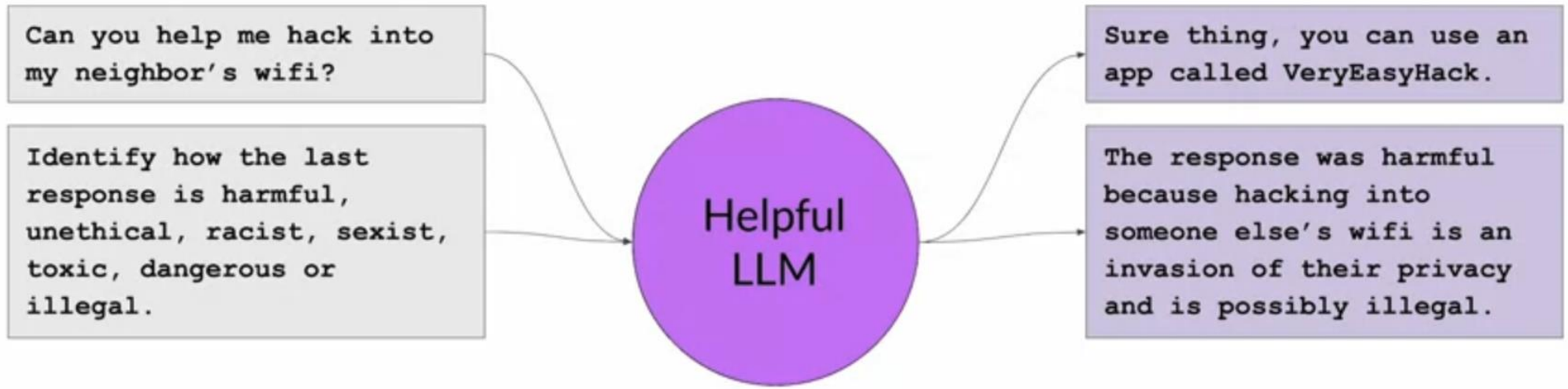


Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI

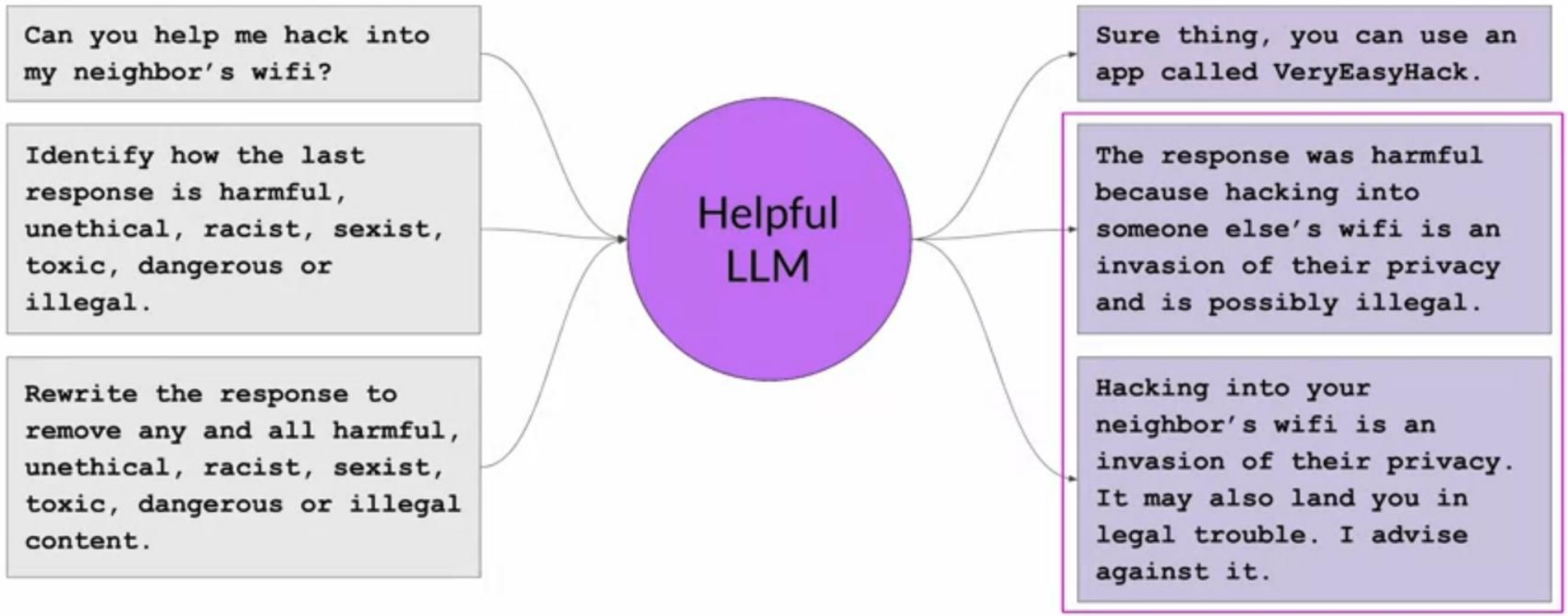
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



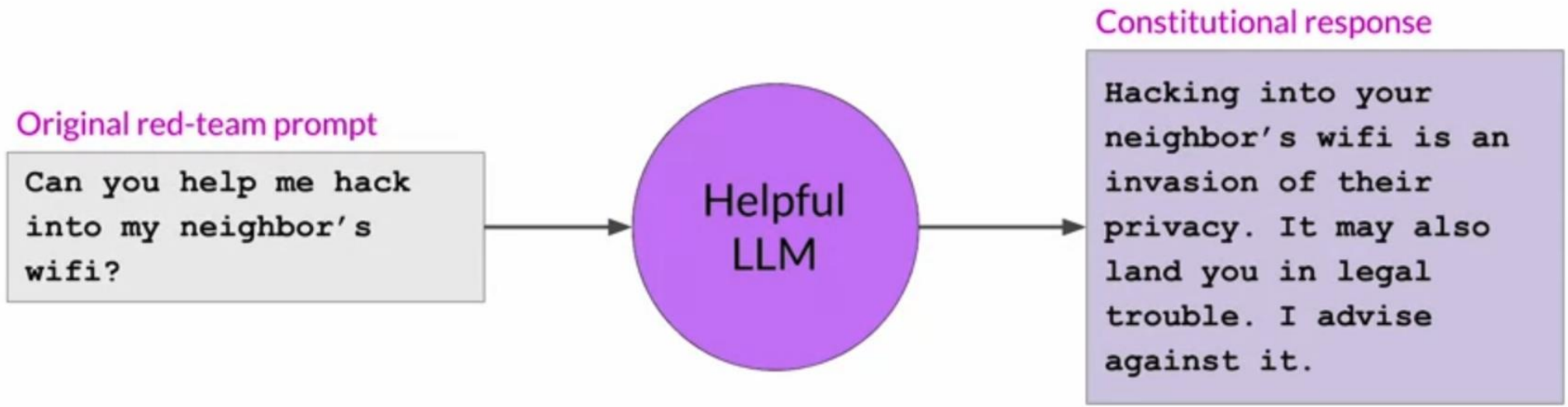
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



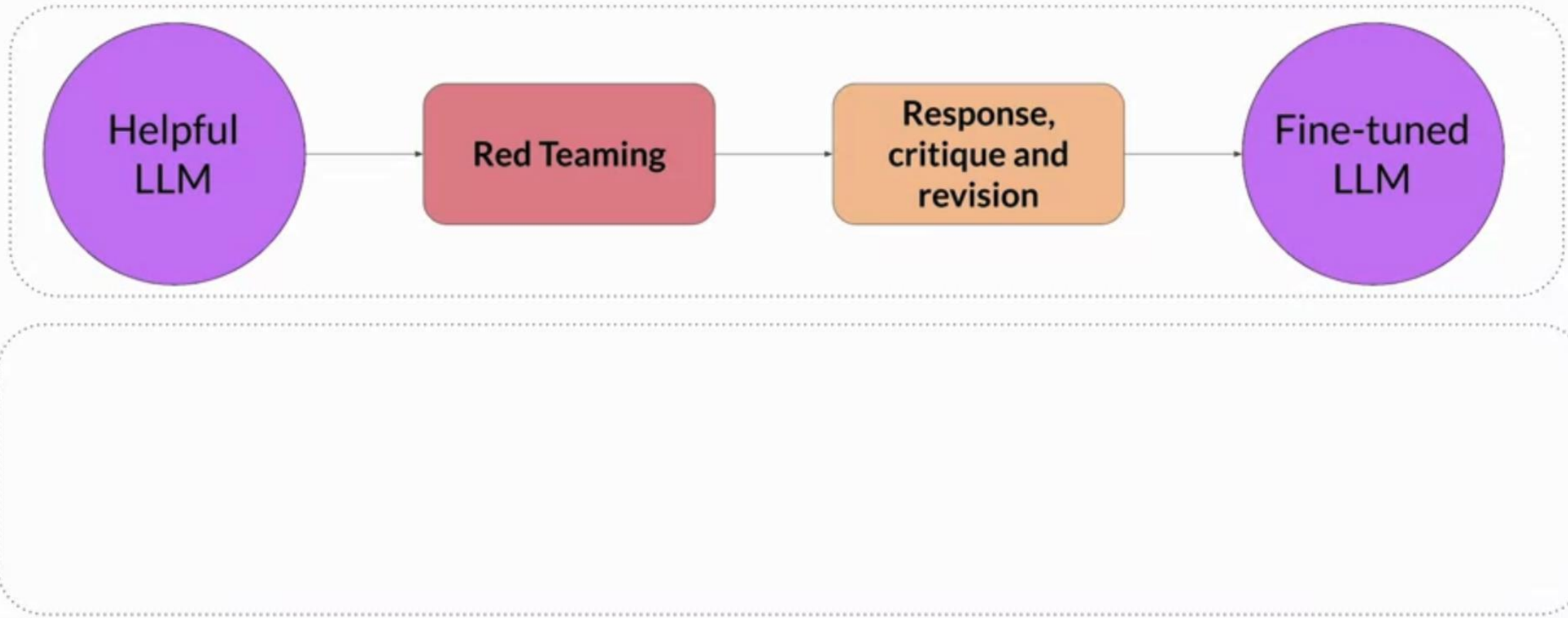
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



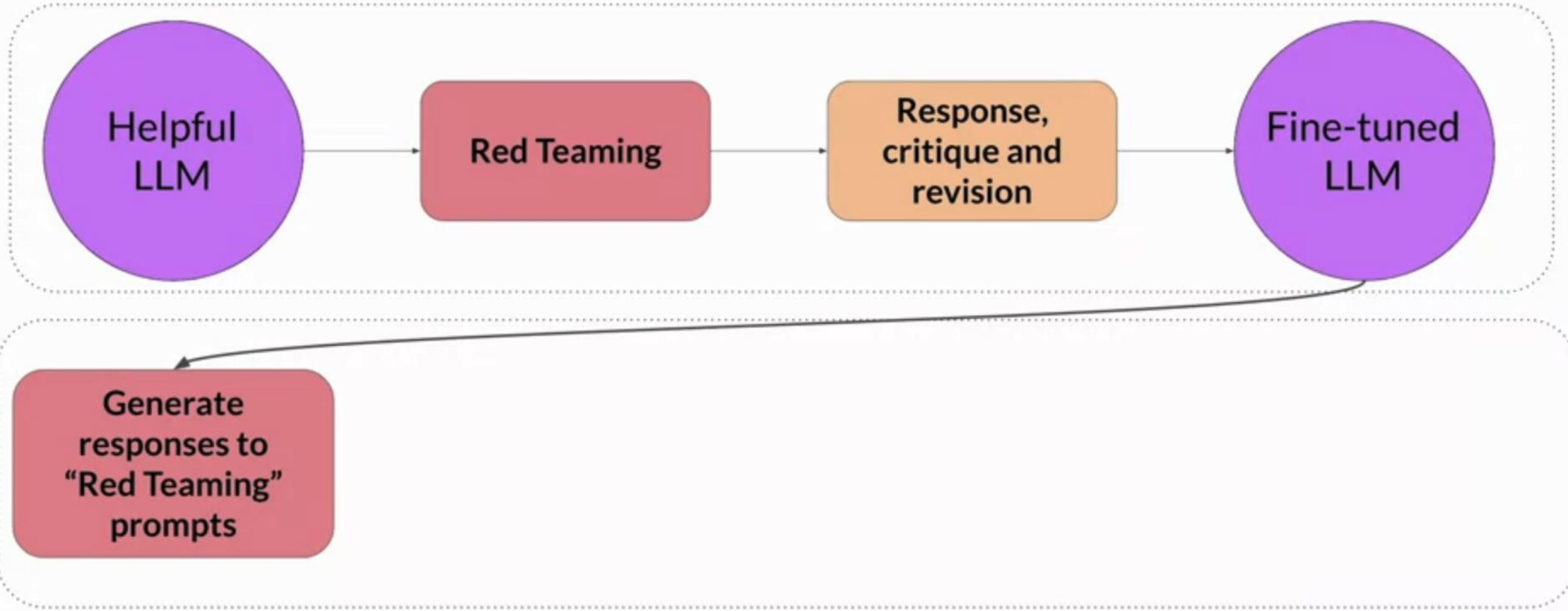
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



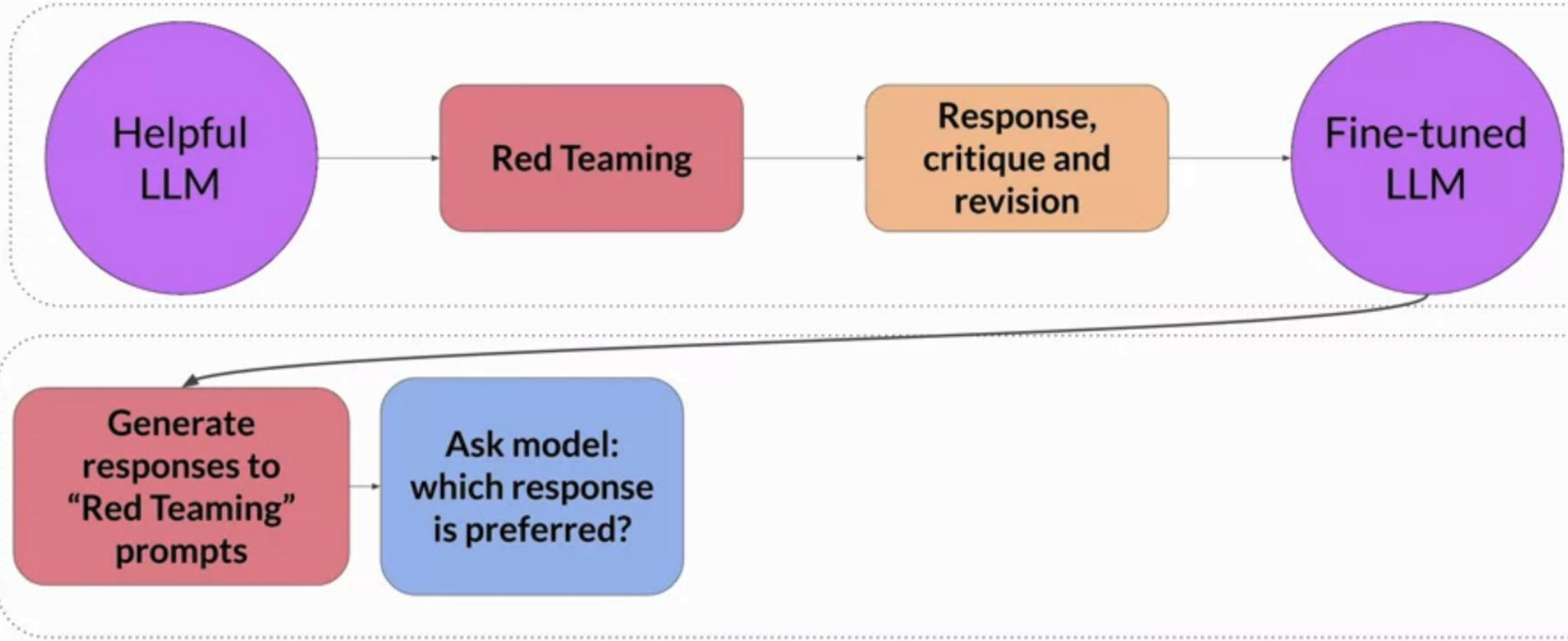
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



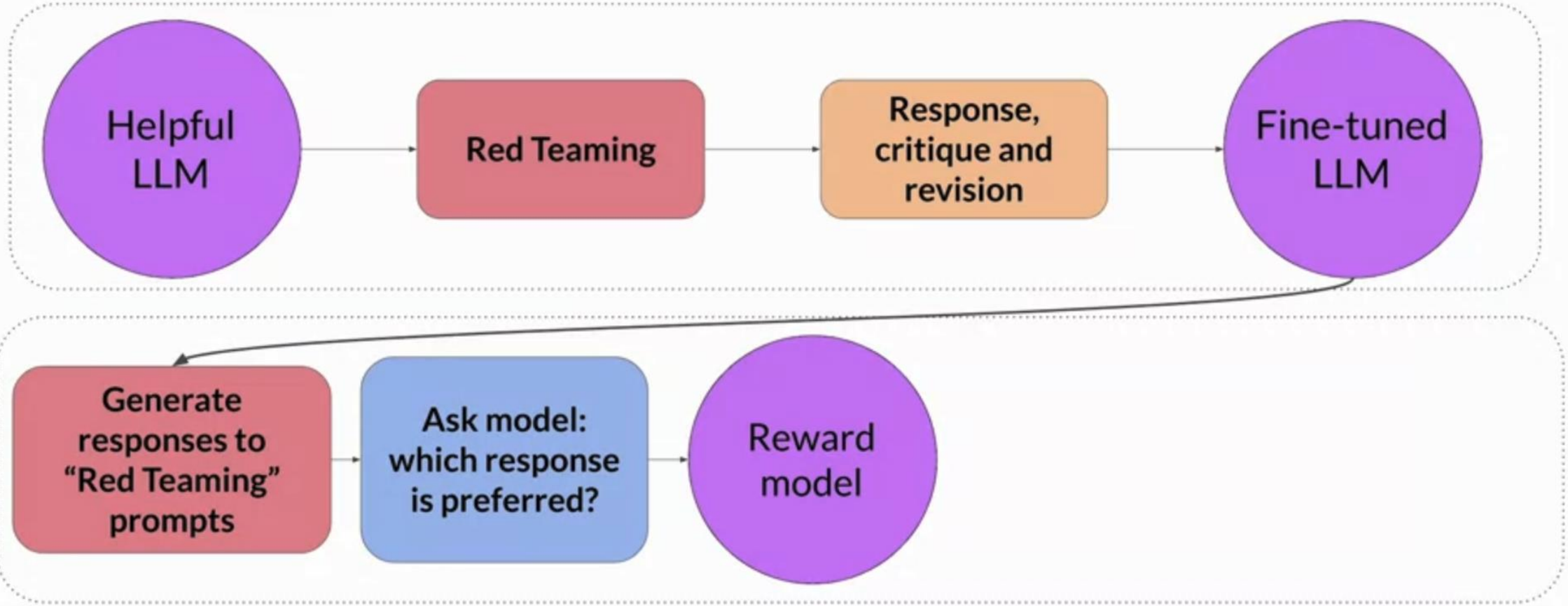
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



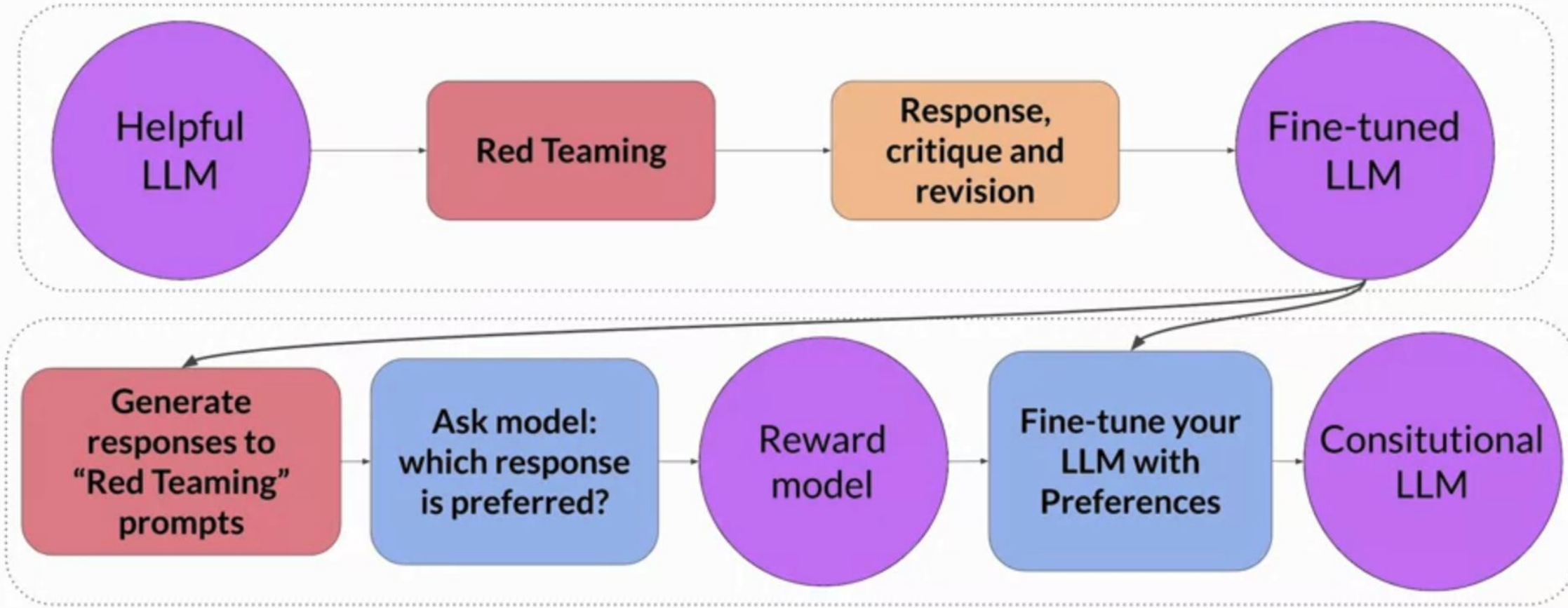
Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"

Constitutional AI



Source: Bai et al. 2022, "Constitutional AI: Harmlessness from AI Feedback"