

# Lecture 20 - Reinforcement Learning from Human Feedback

Guiliang Liu

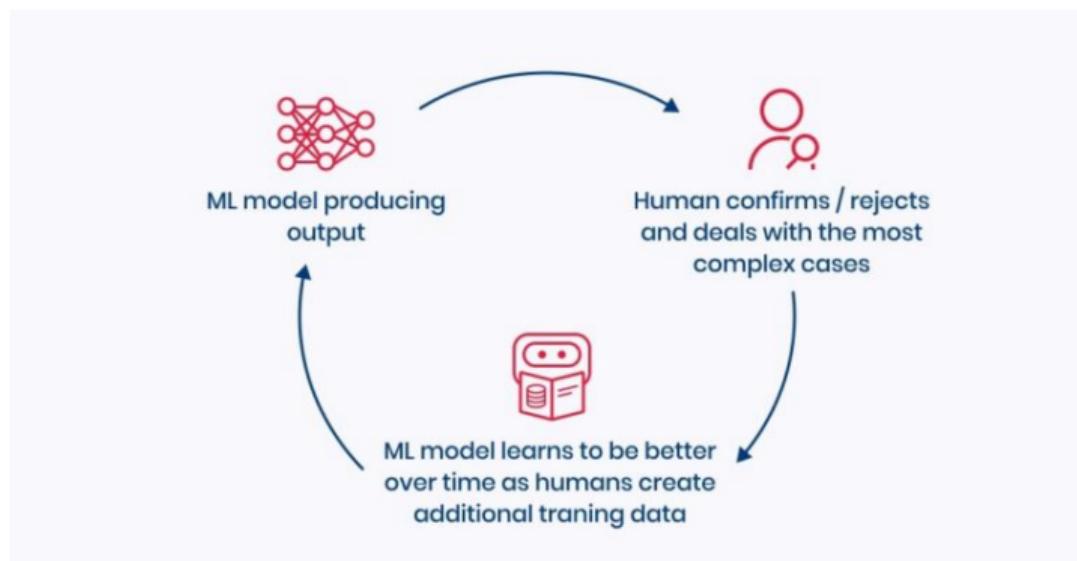
The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning

Course Page: [\[Click\]](#)

# Human in-the-loop Machine Learning

Human-in-the-loop machine learning (HITL) is a branch of artificial intelligence that leverages both human and machine intelligence to create machine learning models.



# Human in-the-loop Machine Learning

Human-in-the-loop machine learning (HITL) is a branch of artificial intelligence that leverages **both human and machine intelligence to create machine learning models.**

- It involves **a continuous feedback loop** where humans train, tune, and test an algorithm, and intervene when the machine is not able to solve a problem.
- HITL improves machine learning over random sampling by selecting **the most critical data needed to refine the model.**



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Human in the Loop Reinforcement Learning

Human-in-the-loop reinforcement learning (HITL-RL) is a form of reinforcement learning where a human interacts directly with the learning process by:

- **Providing feedback:** The human can provide feedback on the agent's actions. This feedback can supplement or replace the rewards in the environment.
- **Setting goals:** The human can define the goals or tasks that the agent should strive to achieve. This can be particularly useful in complex environments where defining a suitable reward function is challenging.
- **Demonstrating /Correcting actions:** The human can show the agent how to perform certain actions or behaviors. If the agent makes a mistake or an incorrect decision, the human can step in and correct it.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Human in the Loop Reinforcement Learning

In this lecture, we are particularly interested in learning reward functions from human feedback, since:

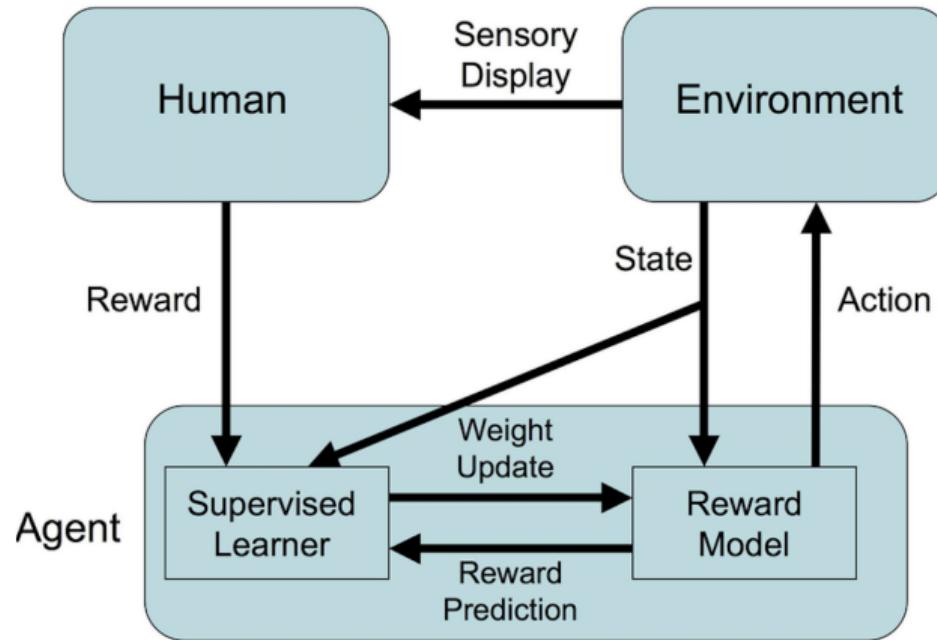
- Reward functions can efficiently represent the signals (including feedback, goals, and preferred actions) listed above.
- In the task where the reward signals are unavailable, the learned reward functions can enable training RL models for efficient control.
- Reward functions summarize human preference, which facilitates the interpretation of reinforcement learning.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Human in the Loop Reinforcement Learning



Knox, W. Bradley, and Peter Stone. "Tamer: Training an agent manually via evaluative reinforcement." 2008 7th IEEE international conference on development and learning. IEEE, 2008.



香港中文大學(深圳)

Hong Kong University of Science and Technology, Shenzhen

# Human in the Loop Reinforcement Learning

However, asking humans to directly design the rewards functions manually is problematic,

- **Sub-optimal Control Performance.** Manual reward engineering might lead to sub-optimal control performance in finishing the task.
- **Safety Issues.** Using a simple proxy for a complex goal, or getting the complex goal a bit wrong, can lead to undesirable and even dangerous behavior.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reinforcement Learning from Human Preferences

Instead of directly providing the reward, humans provide the **ranking of actions**. E.g., in a state  $s$ ,  $a_i > a_j$ , meaning that the  $a_i$  is better than  $a_j$ .



# Reinforcement Learning from Human Preferences

The Plackett-Luce ranking model defines the probability of a state-action pair  $(s, a_i)$  being the largest among a given set  $(s, a_i)_{i=0}^{K-1}$  as:

$$p(a_i > a_j, \forall j \neq i | s) = \frac{\exp(r_\theta(s, a_i))}{\sum_{j=0}^{K-1} \exp(r_\theta(s, a_j))}$$

Let  $\sigma : [K] \rightarrow [K]$  denote the output of the human labeler, which is a permutation function representing the ranking of the actions. The distribution of  $\sigma$  follows:

$$p(\sigma | s, a_0, a_1, \dots, a_{K-1}) = \prod_{k=0}^{K-1} \frac{\exp(r_\theta(s, a_{\sigma(k)}))}{\sum_{j=k}^{K-1} \exp(r_\theta(s, a_{\sigma(j)}))}$$



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reinforcement Learning from Human Preferences

When  $K = 2$ , the above representation reduces to the pairwise comparison of the Bradley-Terry-Luce (BTL) model:

$$p(y = l | s, a_0, a_1) = \frac{\exp(r_\theta(s, a_l))}{\exp(r_\theta(s, a_0)) + \exp(r_\theta(s, a_1))}$$

The MLE objective of RLHF can be defined as:

$$\hat{\theta}_{MLE} \in \arg \max_{\theta} \ell_{\mathcal{D}}(\theta),$$

$$\ell_{\mathcal{D}}(\theta) = - \sum_{i=1}^n \log \left( \sum_{y^i=\{0,1\}} \frac{y^i \cdot \exp r_\theta(s^i, a_{y^i}^i)}{\exp r_\theta(s^i, a_0^i) + \exp r_\theta(s^i, a_1^i)} \right)$$



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reinforcement Learning from Human Preferences

## Lemma

(Strong convexity of  $\ell$ .) We first show that  $\ell_{\mathcal{D}}$  is strongly convex at  $\theta$  with respect to the semi-norm  $\|\cdot\|_{\Sigma_{\mathcal{D}}}$ , meaning that there is some constant  $\gamma > 0$  such that:

$$\ell_{\mathcal{D}}(\theta + \Delta) - \ell_{\mathcal{D}}(\theta) - \langle \nabla \ell_{\mathcal{D}}(\theta^*), \Delta \rangle \leq \gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2$$

for all perturbations  $\Delta \in \mathbb{R}^d$  such that  $\theta + \Delta \in \Theta$ .



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reinforcement Learning from Human Preferences

## Lemma

For any  $\lambda > 0$ , with probability at least  $1 - \delta$ ,

$$\|\hat{\theta}_{MLE} - \theta^*\|_{\Sigma_D + \lambda I} \leq C \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Here  $\Sigma_D = \frac{1}{n} \sum_{i=1}^n (\phi(s^i, a_1^i) - \phi(s^i, a_0^i))(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))^T$ ,  
 $\gamma = 1/(2 + \exp(-LB) + \exp(LB))$ .



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reinforcement Learning from Human Preferences

Instead of labeling state-action pairs, one can label the entire trajectory. The MLE objective of RLHF can be updated to:

$$\hat{\theta}_{MLE} \in \arg \max_{\theta} \ell_{\mathcal{D}}(\theta),$$

$$\ell_{\mathcal{D}}(\theta) = - \sum_{i=1}^n \log \left( \sum_{y^i=\{0,1\}} \frac{y^i \cdot \exp \sum_t r_{\theta}(s_t^i, a_{t,y^i}^i)}{\exp \sum_t r_{\theta}(s_t^i, a_{0,t}^i) + \exp \sum_t r_{\theta}(s_t^i, a_{1,t}^i)} \right)$$

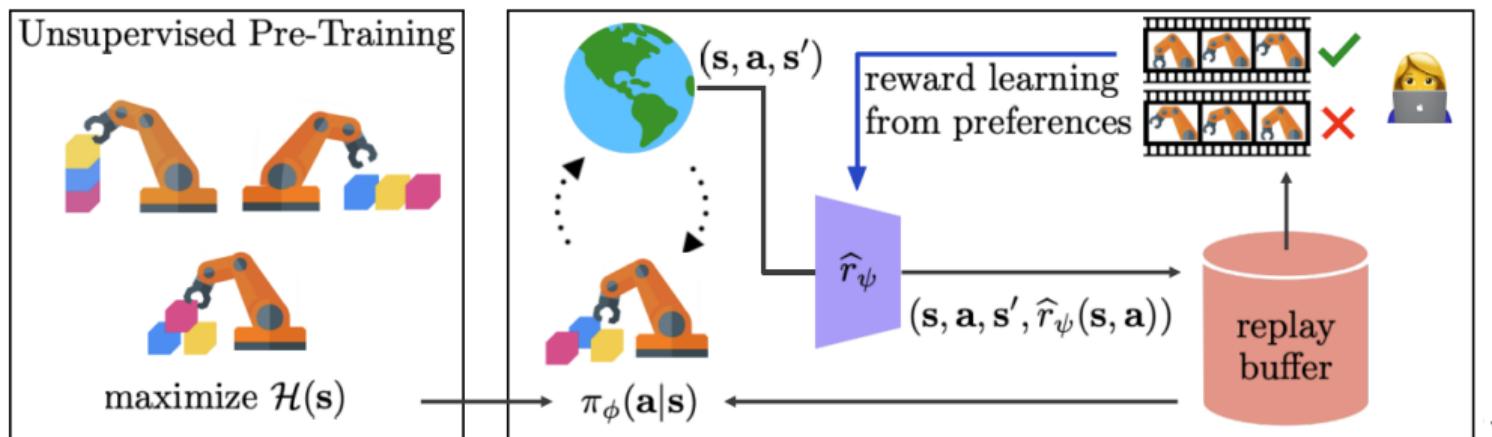


香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

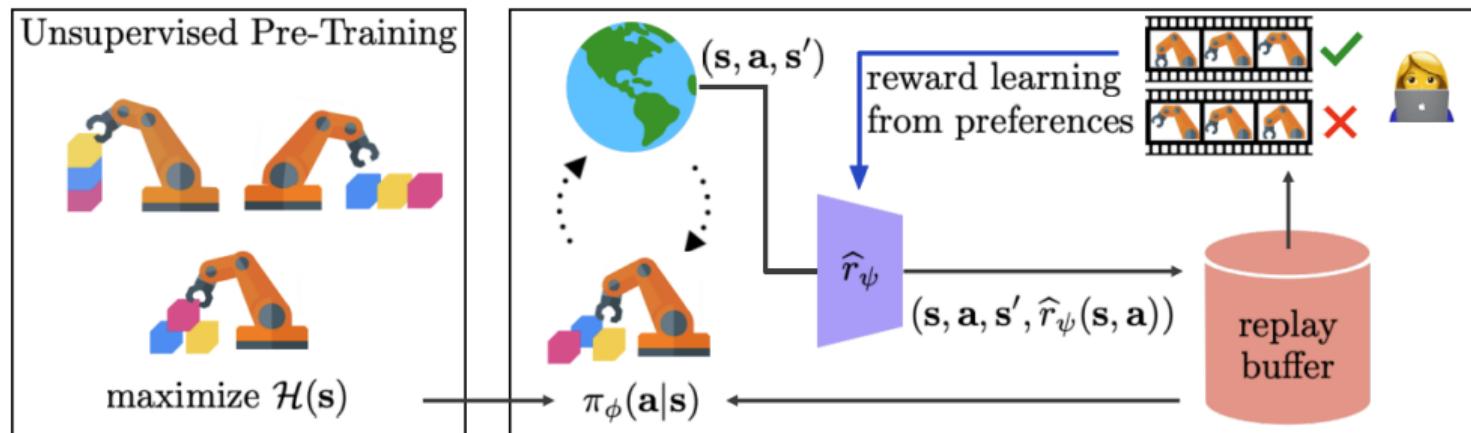
# Reinforcement Learning from Human Preferences

**Preference-Based Reward Learning.** Step 1) The agent engages in unsupervised pre-training during which it is encouraged to visit a diverse set of states to collect diverse experiences (left).



# Reinforcement Learning from Human Preferences

**Preference-Based Reward Learning.** Step 2) A teacher provides preferences between two clips of behavior. A reward model is learned based on them. The agent is updated to maximize the expected return under the model (right).



# Train ChatGPT with RLHF

## What is GPT?

Generative Pre-trained Transformers (GPT), commonly known as GPT, are a family of neural network models that [use the transformer architecture](#) and are a key advancement in artificial intelligence (AI) powering generative AI applications such as ChatGPT.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Train ChatGPT with RLHF

Trained a GPT model using Reinforcement Learning from Human Feedback (RLHF):

- Train an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant.
- Learn a reward model. Collect comparison data, which consisted of two or more model responses ranked by quality by 1) randomly selecting a model-written message, 2) sampling several alternative completions, and 3) having AI trainers rank them. Learn the model based on the collected dataset.
- Fine-tune the GPT model using Proximal Policy Optimization (PPO).



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Train ChatGPT with RLHF

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain reinforcement learning to a 6 year old.



We give treats and punishments to teach...

A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

Explain reinforcement learning to a 6 year old.



D > C > A > B



D > C > A > B

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 the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

Write a story about otters.



Once upon a time...



$r_k$

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

深圳)  
Hong Kong, Shenzhen

# Train a GPT Model using Supervised Fine-Tuning

GPT models are pre-trained over a corpus/dataset of unlabeled textual data using a language modeling objective by

- sampling some text from the dataset.
- training the model to predict the next word.

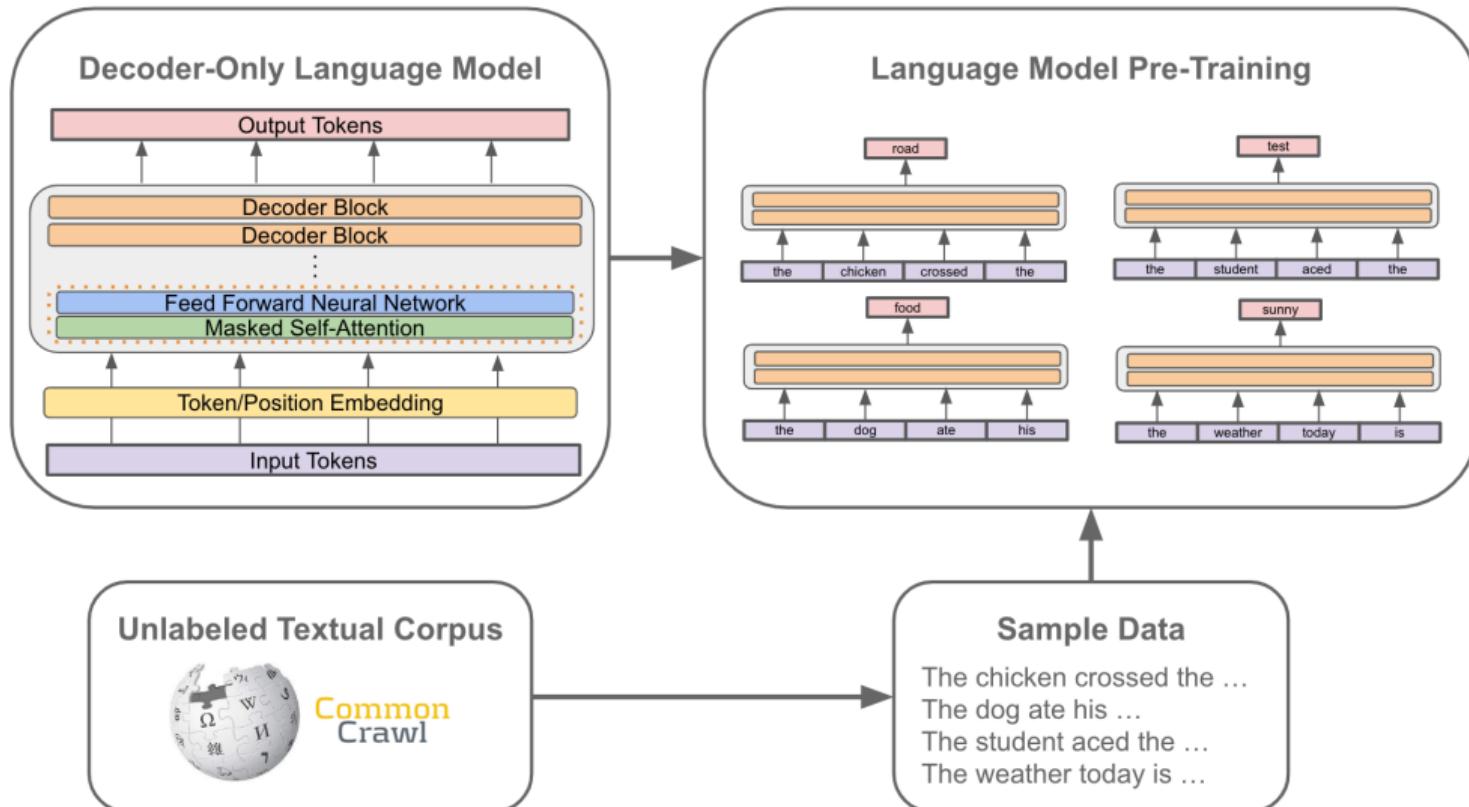
This pre-training procedure is a form of self-supervised learning, as the correct "next" word can be determined by simply looking at the next word in the dataset.



香港中文大學(深圳)

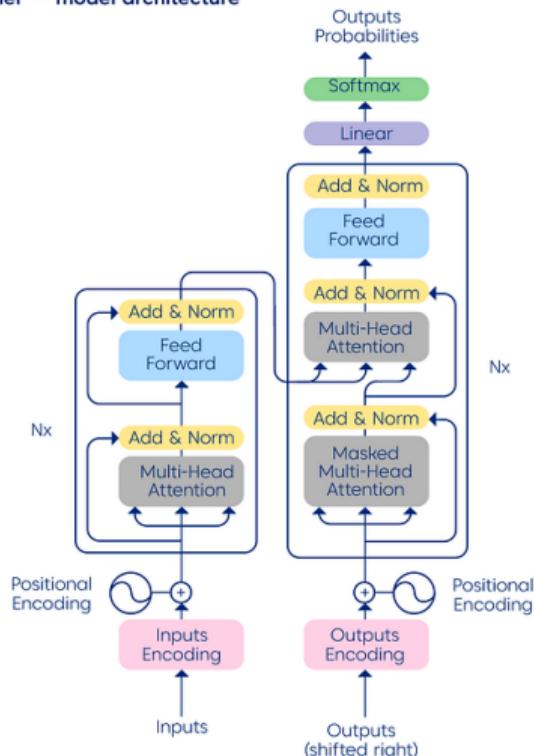
The Chinese University of Hong Kong, Shenzhen

# Train a GPT Model using Supervised Fine-Tuning



# Train a GPT Model using Supervised Fine-Tuning

The Transformer — model architecture



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen

# Train a GPT Model using Supervised Fine-Tuning

Let us denote this set of tokens (of size  $N$ ) that comprise our pre-training dataset:

$$\mu = \{\mu_1, \mu_2, \dots, \mu_N\}$$

Given a deep learning model with parameters  $\theta$ , a language modeling objective tries to maximize the likelihood shown below.

$$\ell(\theta) = \sum_{i=1}^N \log(P(\mu_i | \mu_{i-k}, \dots, \mu_{i-1}; \theta))$$

This loss characterizes the model's probability of predicting the correct next token given  $k$  preceding tokens as context



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Train a GPT Model using Supervised Fine-Tuning

GPT uses a **decoder-only** transformer architecture. A decoder-only architecture removes the following components from the transformer:

- The entire encoder module;
- All encoder-decoder self-attention modules in the decoder.

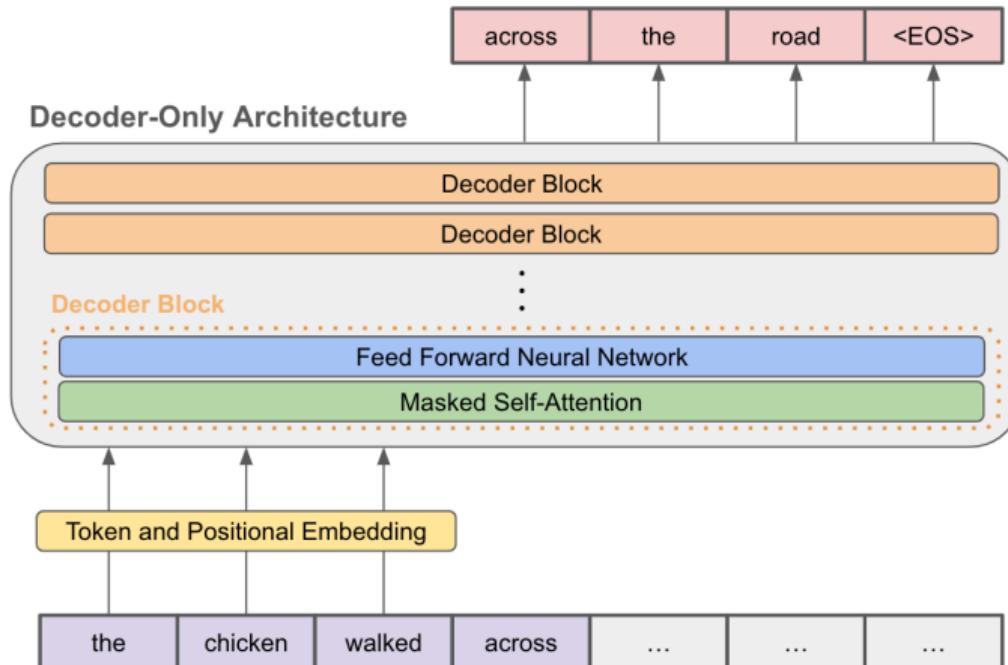
After these components have been removed, each layer of the decoder simply consists of a masked self-attention layer followed by a feed-forward neural network. Stacking several of such layers on top of each other forms a deep, decoder-only transformer architecture.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Train a GPT Model using Supervised Fine-Tuning



# 文大學(深圳) University of Hong Kong, Shenzhen

# Learn a Reward Model

Collect comparison data. Below is a screenshot of the UI that OpenAI's labelers used to create training data for InstructGPT's RM. Labelers rank the responses in the order of preference, but only the ranking is used to train the RM.

The screenshot shows a web-based labeling interface. At the top, there are buttons for 'Submit' and 'Skip'. To the right, it displays 'Page 3 / 11' and 'Total time: 05:39'. The main area is divided into sections: 'Instruction' and 'Output A'. Under 'Instruction', there is a task description: 'Summarize the following news article:' followed by four lines of placeholder text: '==== {article} ====='. Under 'Output A', there is a summary: 'summary1'. Below this, there is a rating scale: 'Rating (1 = worst, 7 = best)' with a horizontal row of numbered boxes from 1 to 7. The number 1 is highlighted with a light blue background. At the bottom, there is a section for notes: 'Notes' with a text input field containing '(Optional) notes'. On the right side of the interface, there is a vertical sidebar with the text '文大學(深圳)' and 'The Chinese University of Hong Kong, Shenzhen'.

Submit Skip Page 3 / 11 Total time: 05:39

Instruction

Output A

Include output

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

文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen

# Learn a Reward Model

Collect comparison data. The inter-labeler agreement is around 73%, which means if they ask 10 people to rank 2 responses, 7 of them will have the same ranking.

**Ranking outputs**

To be ranked

| Rank 1 (best)   | Rank 2   | Rank 3  | Rank 4   | Rank 5 (worst) |
|---|--|---|--|----------------|
| <b>B</b> A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations... | <b>C</b> Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups... | <b>E</b> Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds. | <b>D</b> Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability |                |

# Learn a Reward Model

Learn the reward model. Given a set of tokens (of size N):

$$\mu = [\mu_1, \mu_2, \dots, \mu_N]$$

Let  $\mu_q$  denotes the question and  $\mu_r$  denotes the response. Let's denote

$s_t = [\mu_q, \mu_{r,1}, \mu_{r,2}, \dots, \mu_{r,t-1}]$  ( $t \leq N$ ) and action  $a_t = \mu_{r,t}$ . Learning the reward model  $r_\theta(\cdot)$  (shares the similar structure as the actor) by utilizing the MLE loss:

$$\hat{\theta}_{MLE} \in \arg \max_{\theta} \ell_{\mathcal{D}}(\theta),$$

$$\ell_{\mathcal{D}}(\theta) = - \sum_{i=1}^n \log \left( \sum_{y^i=\{0,1\}} \frac{y^i \cdot \exp \sum_t r_\theta(s_t^i, a_{t,y^i}^i)}{\exp \sum_t r_\theta(s^i, a_{0,t}^i) + \exp \sum_t r_\theta(s_t^i, a_{1,t}^i)} \right)$$



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Fine-tune the GPT model

Fine-tune the GPT model with PPO:

---

**Algorithm 1** PPO-Clip

---

- 1: Input: initial policy parameters  $\theta_0$ , initial value function parameters  $\phi_0$
- 2: **for**  $k = 0, 1, 2, \dots$  **do**
- 3:   Collect set of trajectories  $\mathcal{D}_k = \{\tau_i\}$  by running policy  $\pi_k = \pi(\theta_k)$  in the environment.
- 4:   Compute rewards-to-go  $\hat{R}_t$ .
- 5:   Compute advantage estimates,  $\hat{A}_t$  (using any method of advantage estimation) based on the current value function  $V_{\phi_k}$ .
- 6:   Update the policy by maximizing the PPO-Clip objective:

$$\theta_{k+1} = \arg \max_{\theta} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \min \left( \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_k}(a_t | s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right),$$

typically via stochastic gradient ascent with Adam.

- 7:   Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \left( V_{\phi}(s_t) - \hat{R}_t \right)^2,$$



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

typically via some gradient descent algorithm.

- 8: **end for**

# Limitation of ChatGPT

ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as:

- During RL training, there's currently no source of truth;
- Training the model to be more cautious causes it to decline questions that it can answer correctly;
- Supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Reference

- RLHF: Reinforcement Learning from Human Feedback  
<https://huyenchip.com/2023/05/02/rlhf.html>.
- Introducing ChatGPT. <https://openai.com/blog/chatgpt>
- Language Models: GPT and GPT-2.  
<https://cameronrwolfe.substack.com/p/language-models-gpt-and-gpt-2>.
- Zhu, Banghua, Jiantao Jiao, and Michael I. Jordan. "Principled Reinforcement Learning with Human Feedback from Pairwise or  $K$ -wise Comparisons." arXiv preprint arXiv:2301.11270 (2023).
- Lee, Kimin, Laura M. Smith, and Pieter Abbeel. "PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training." International Conference on Machine Learning. PMLR, 2021.



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen

# Question and Answering (Q&A)



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen