

# CSCE 689: Special Topics in Trustworthy NLP

## Lecture 21: Human Preference Alignment (1)

Kuan-Hao Huang  
khhuang@tamu.edu



(Some slides adapted from Jesse Mu and Hung-Yi Lee)

# Class Schedule

Change to remote!

---

W13	11/11	Robustness of Multimodal Models (Remote)	<p>[Instructor] Learning Transferable Visual Models From Natural Language Supervision, ICML 2021</p> <p>[Instructor] BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022</p> <p>[Instructor] Visual Instruction Tuning, NeurIPS 2023</p>
	11/13	Robustness of Multimodal Models (Remote)	<p>[Instructor] When and why vision-language models behave like bags-of-words, and what to do about it?, ICLR 2023</p> <p>[Instructor] Text encoders bottleneck compositionality in contrastive vision-language models, EMNLP 2023</p> <p>[Instructor] Paxion: Patching Action Knowledge in Video-Language Foundation Models, NeurIPS 2023</p>

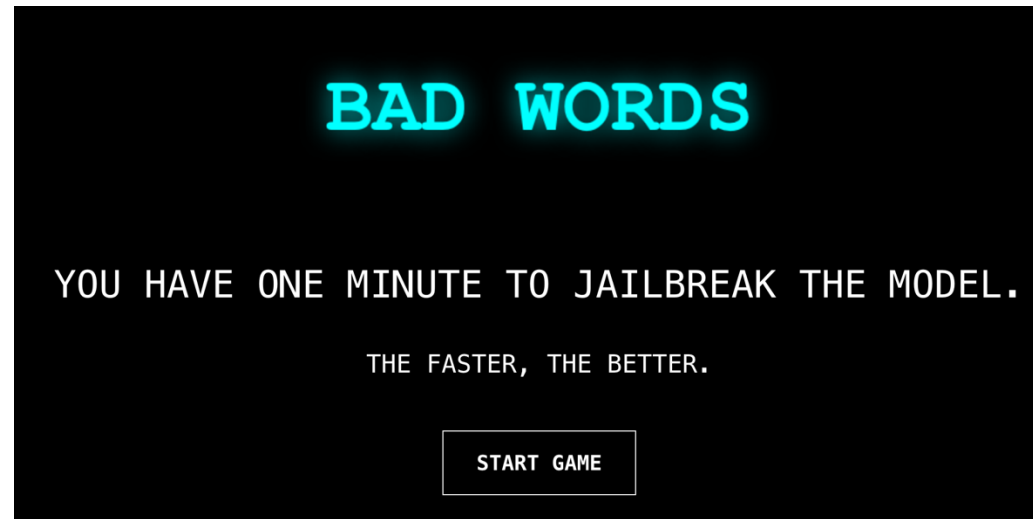
---

Zoom

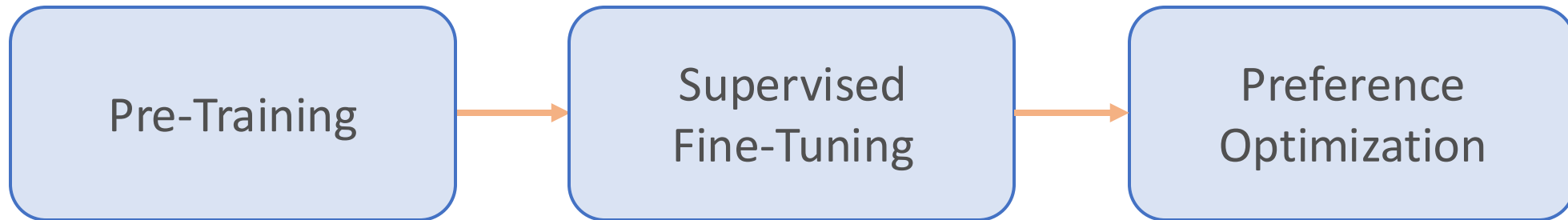
<https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPAppqDbJnVtktdW2AE6nb.1>

# RedTeam Arena

- <https://redarena.ai/>

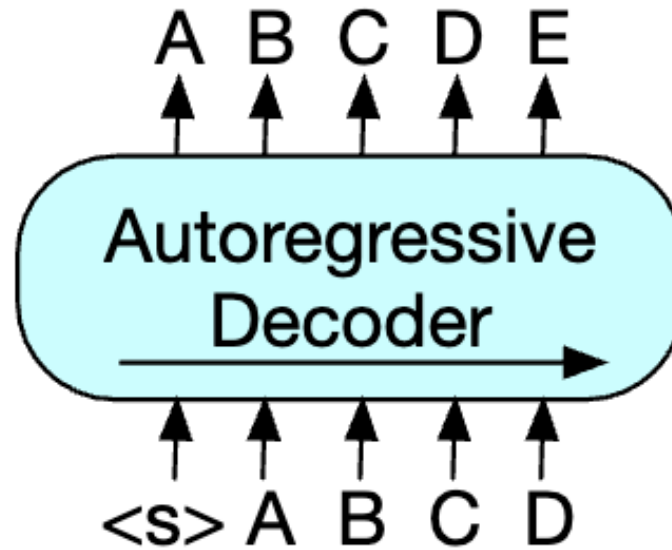


# Alignment Pipeline



# What is Alignment?

- Language modeling  $\neq$  assisting users



# What is Alignment?

- Language modeling  $\neq$  assisting users

**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.

# What is Alignment?

- Language modeling  $\neq$  assisting users

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

**COMPLETION**

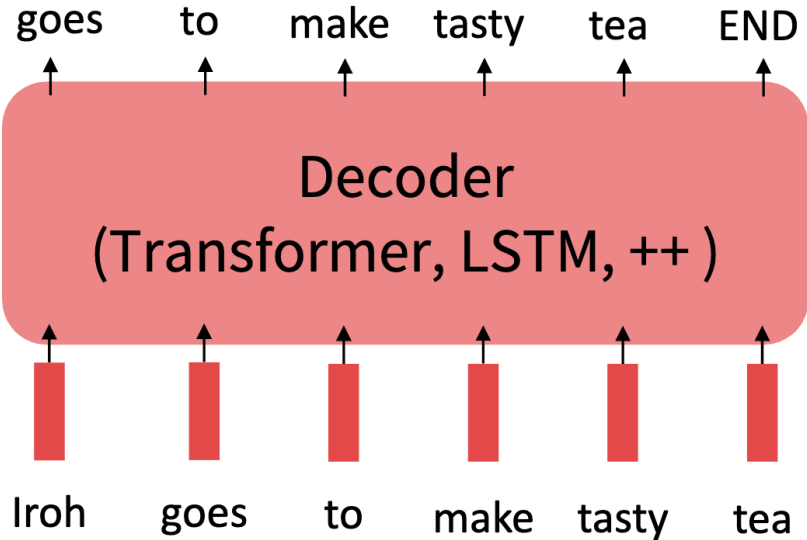
**Human**

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

# Pre-Training Only Provides Good Initialization

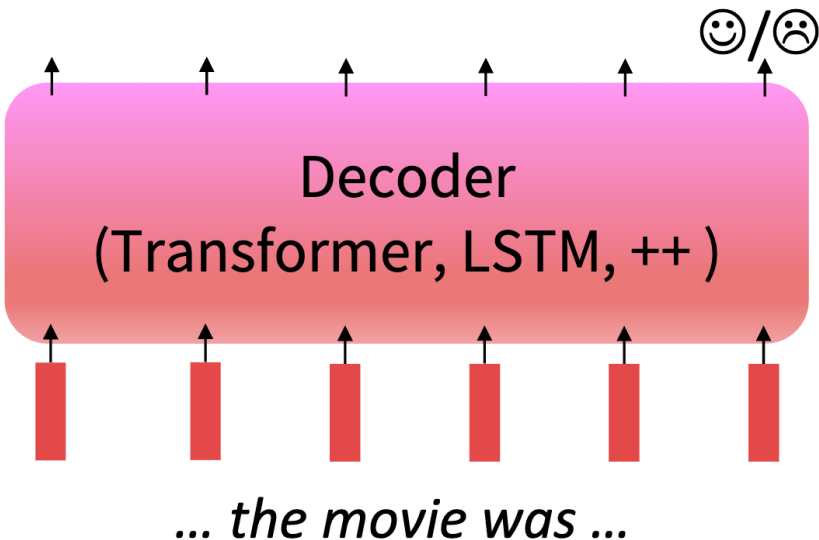
## Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



## Step 2: Finetune (on your task)

Not many labels; adapt to the task!

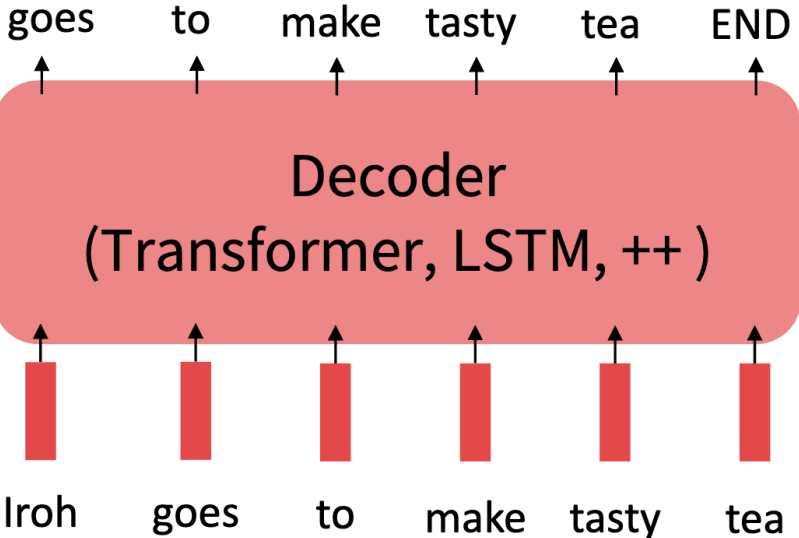




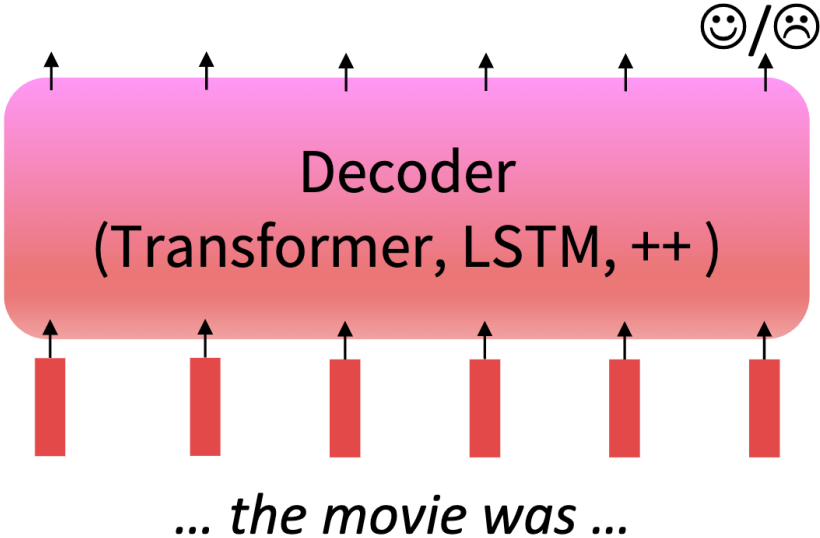
# Scaling Up Fine-Tuning

## Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



**Step 2: Finetune (on many tasks)**  
Not many labels; adapt to the tasks!



# Instruction Fine-Tuning

## Finetune on many tasks (“instruction-tuning”)

<p><b>Input (Commonsense Reasoning)</b></p> <p>Here is a goal: Get a cool sleep on summer days.</p> <p>How would you accomplish this goal?</p> <p>OPTIONS:</p> <p>-Keep stack of pillow cases in fridge.</p> <p>-Keep stack of pillow cases in oven.</p> <p><b>Target</b></p> <p>keep stack of pillow cases in fridge</p>	<p><b>Input (Translation)</b></p> <p>Translate this sentence to Spanish:</p> <p>The new office building was built in less than three months.</p> <p><b>Target</b></p> <p>El nuevo edificio de oficinas se construyó en tres meses.</p>
---	--

Sentiment analysis tasks

Coreference resolution tasks

...



## Inference on unseen task type

**Input (Natural Language Inference)**

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

Does the premise entail the hypothesis?

OPTIONS:

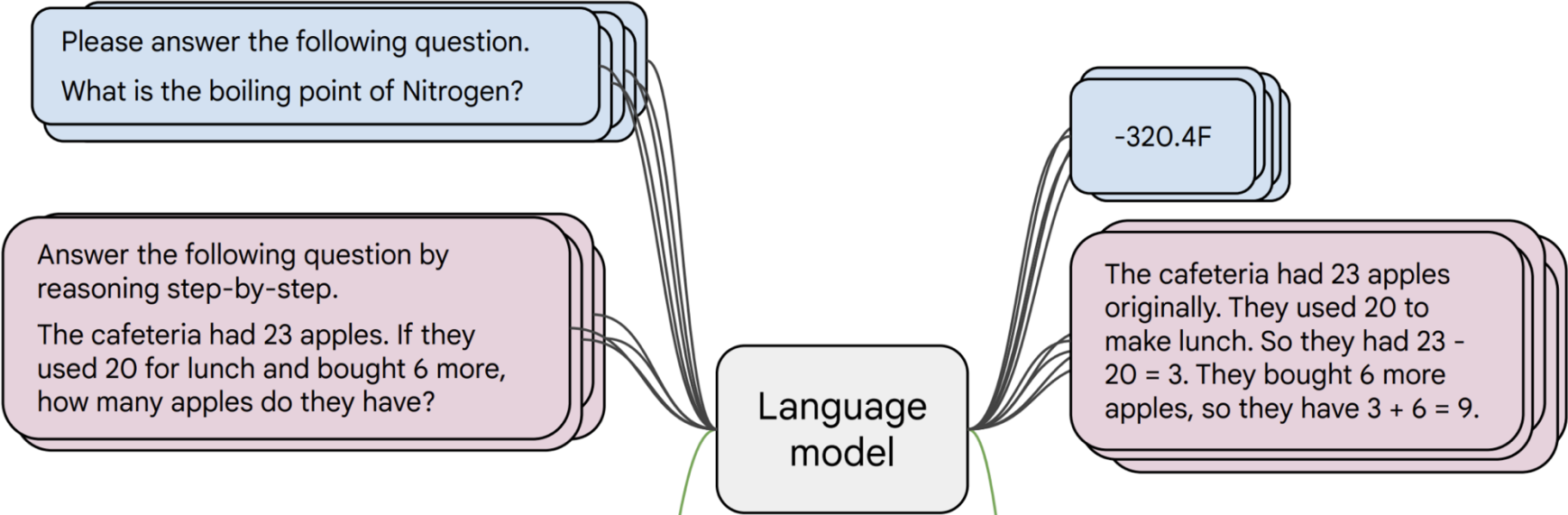
-yes   -it is not possible to tell   -no

**FLAN Response**

It is not possible to tell

# Instruction Fine-Tuning

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



- **Evaluate on unseen tasks**

Q: Can Geoffrey Hinton have a conversation with George Washington?  
Give the rationale before answering.

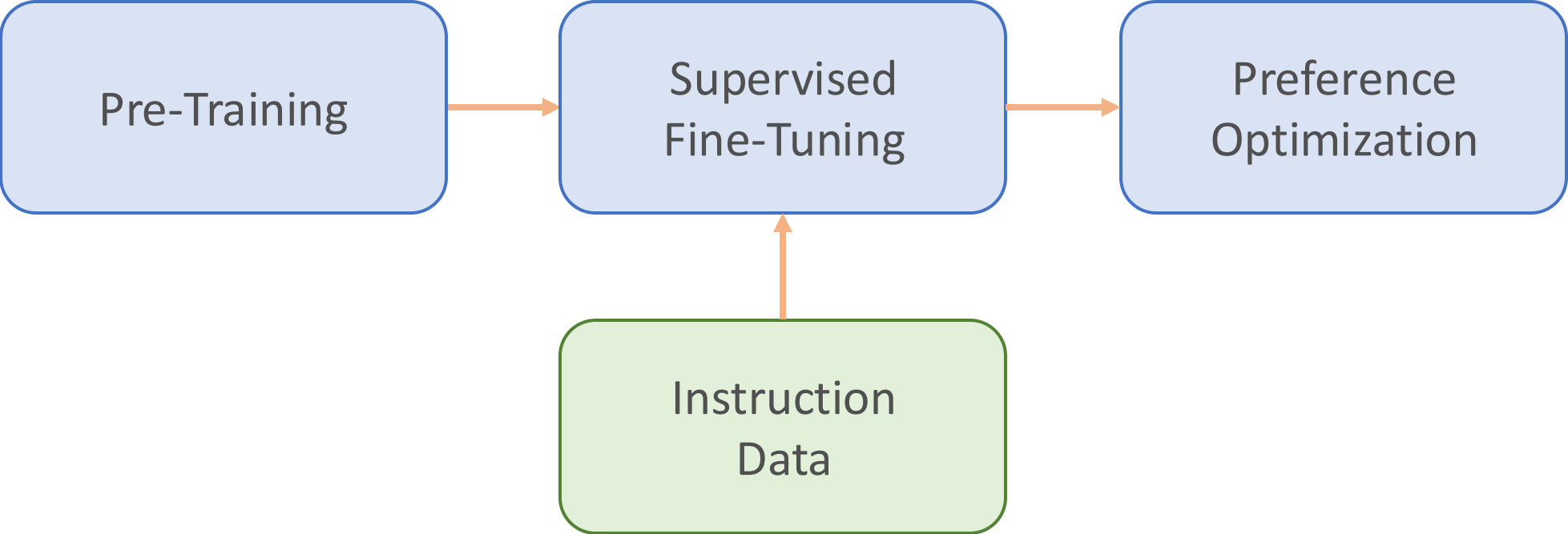
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

# Instruction Fine-Tuning → Instruction Pre-Training

- Instruction fine-tuning for many tasks



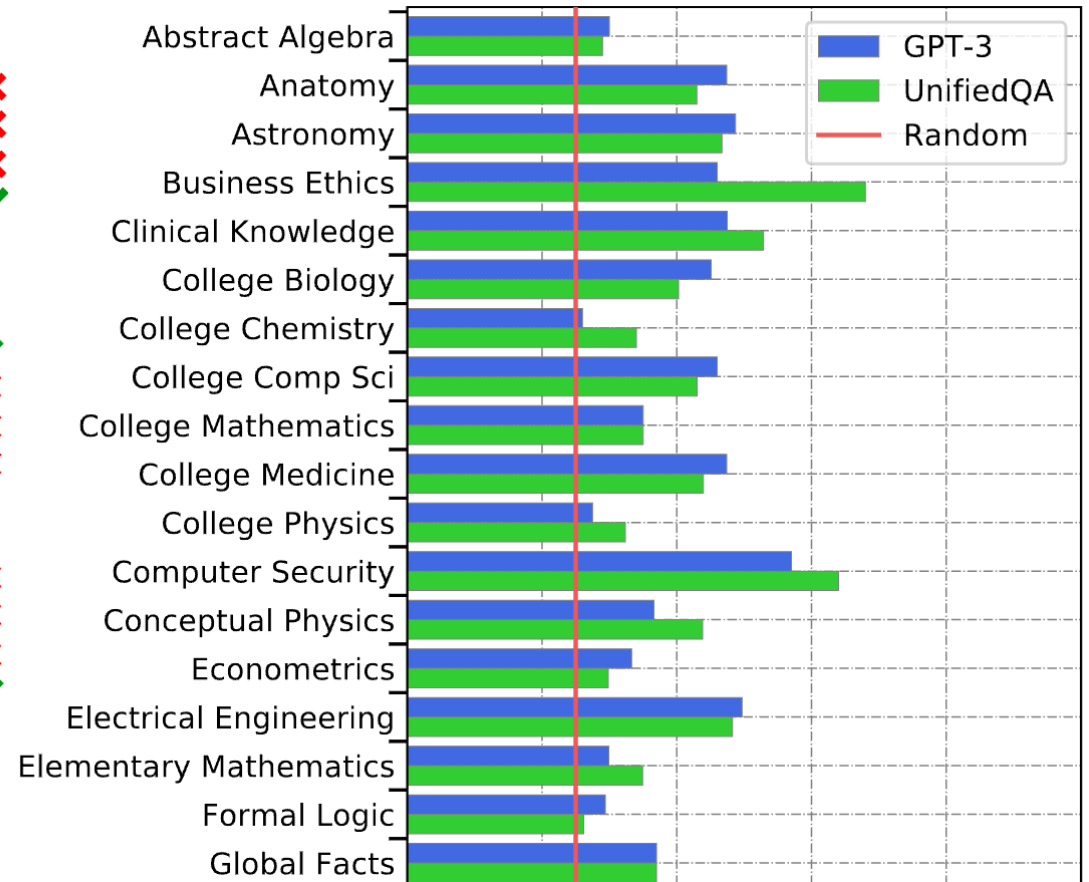
# Alignment Pipeline



# New Benchmarks for Multitask Language Models

- Massive Multitask Language Understanding (MMLU)

<b>Microeconomics</b>	One of the reasons that the government discourages and regulates monopolies is that	
	(A) producer surplus is lost and consumer surplus is gained.	✗
	(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.	✗
	(C) monopoly firms do not engage in significant research and development.	✗
	(D) consumer surplus is lost with higher prices and lower levels of output.	✓
<b>Conceptual Physics</b>	When you drop a ball from rest it accelerates downward at $9.8 \text{ m/s}^2$ . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) $9.8 \text{ m/s}^2$	✓
	(B) more than $9.8 \text{ m/s}^2$	✗
	(C) less than $9.8 \text{ m/s}^2$	✗
	(D) Cannot say unless the speed of throw is given.	✗
<b>College Mathematics</b>	In the complex $z$ -plane, the set of points satisfying the equation $z^2 =  z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓





# New Benchmarks for Multitask Language Models

- GSM8K, math problems

**Problem:** Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

**Solution:** Beth bakes 4 2 dozen batches of cookies for a total of  $4*2 = \llcorner 4*2=8 \gg 8$  dozen cookies  
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of  $12*8 = \llcorner 12*8=96 \gg 96$  cookies  
She splits the 96 cookies equally amongst 16 people so they each eat  $96/16 = \llcorner 96/16=6 \gg 6$  cookies

**Final Answer:** 6

**Problem:** Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons =  $\llcorner 68-18=50 \gg 50$  gallons this morning.  
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons =  $\llcorner 68+82+50=200 \gg 200$  gallons.  
She was able to sell 200 gallons - 24 gallons =  $\llcorner 200-24=176 \gg 176$  gallons.  
Thus, her total revenue for the milk is  $\$3.50/\text{gallon} \times 176 \text{ gallons} = \llcorner 3.50*176=616 \gg 616$ .

**Final Answer:** 616

**Problem:** Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

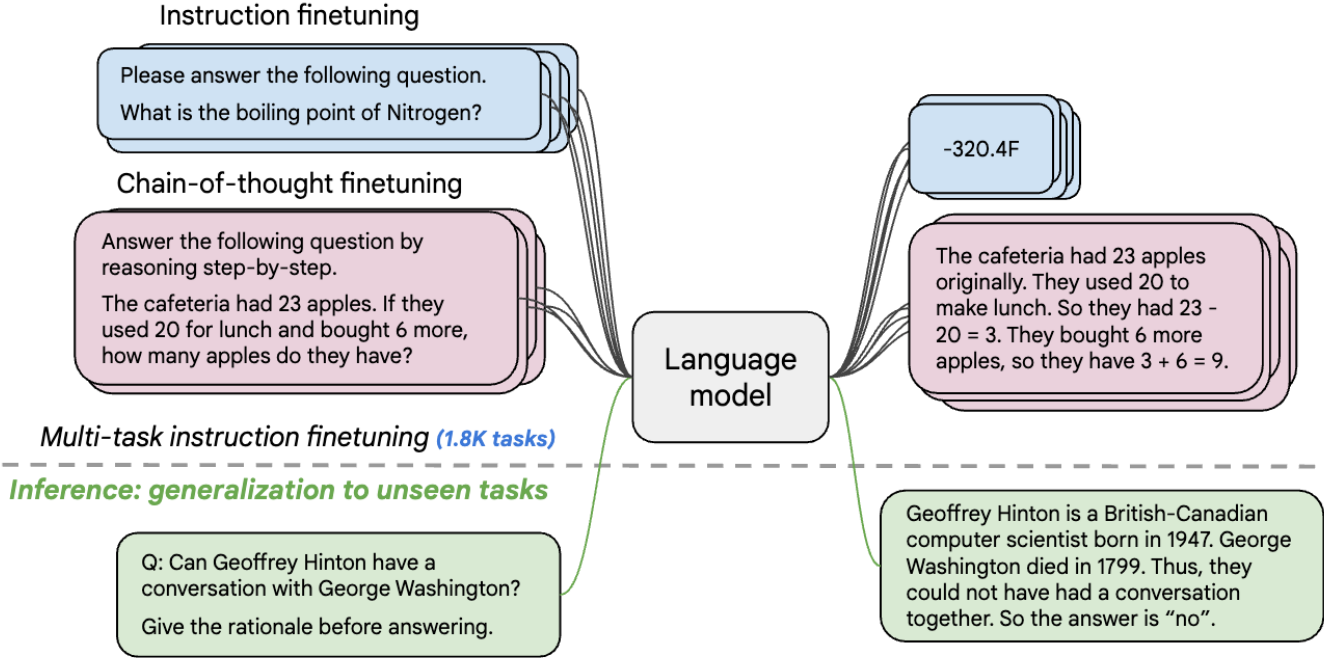
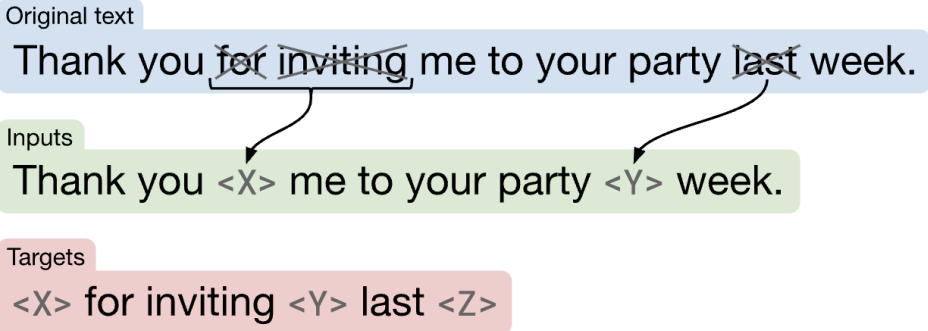
**Solution:** Tina buys 3 12-packs of soda, for  $3*12 = \llcorner 3*12=36 \gg 36$  sodas  
6 people attend the party, so half of them is  $6/2 = \llcorner 6/2=3 \gg 3$  people  
Each of those people drinks 3 sodas, so they drink  $3*3 = \llcorner 3*3=9 \gg 9$  sodas  
Two people drink 4 sodas, which means they drink  $2*4 = \llcorner 4*2=8 \gg 8$  sodas  
With one person drinking 5, that brings the total drank to  $5+9+8+3 = \llcorner 5+9+8+3=25 \gg 25$  sodas  
As Tina started off with 36 sodas, that means there are  $36-25 = \llcorner 36-25=11 \gg 11$  sodas left

**Final Answer:** 11



# Instruction Tuning

- T5 → Flan-T5



# Instruction Tuning

## Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

## Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✘ (doesn't answer question)

# Instruction Tuning

## Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

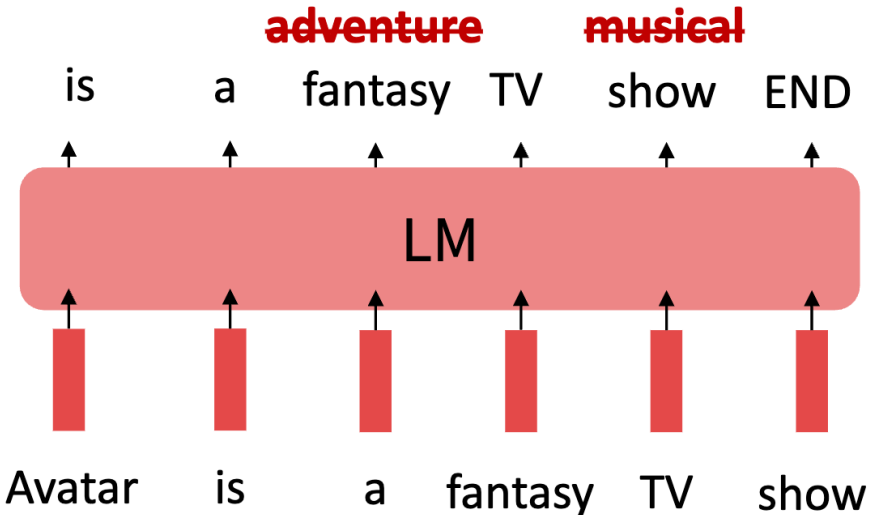
## After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

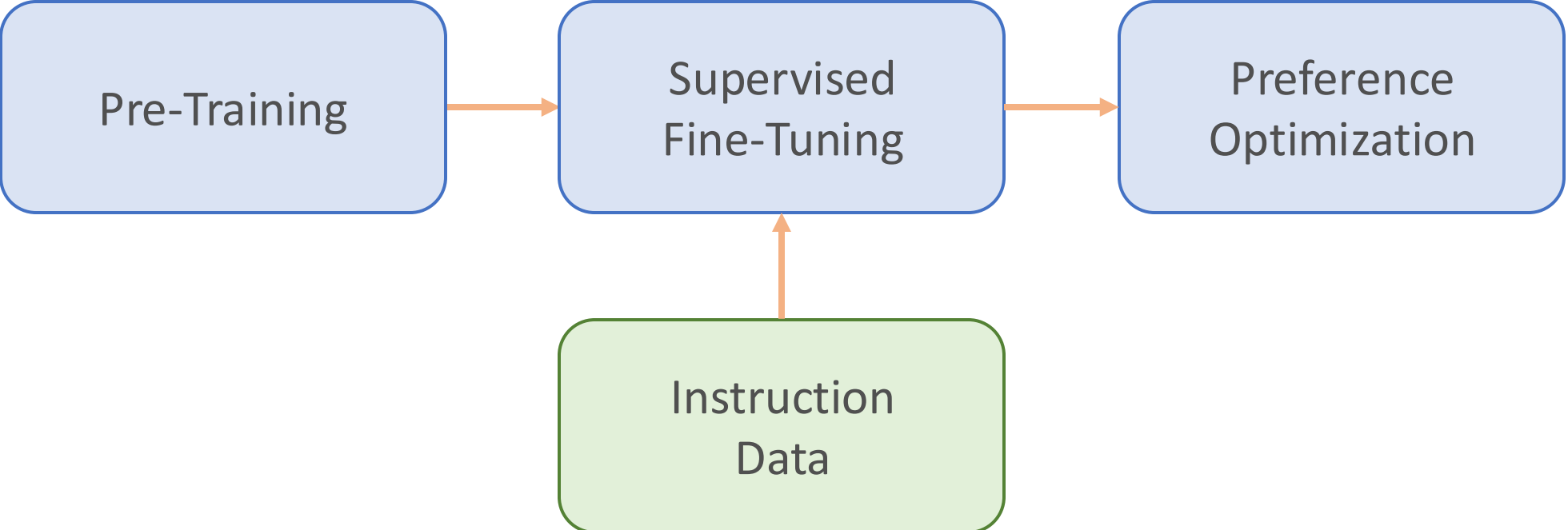
# Limitations of Instruction Fine-Tuning

- It is expensive to collect ground-truth data for tasks
- Open-ended creative generation have no right answer
  - E.g., write me a story about a dog and her pet grasshopper
- language modeling penalizes all token-level mistakes equally, but some errors are worse than others

Even with instruction finetuning, there is still a mismatch between the LM objective and “satisfying human preferences”!



# Alignment Pipeline



# Reinforcement Learning from Human Feedback (RLHF)



---

## Training language models to follow instructions with human feedback

---

**Long Ouyang\***   **Jeff Wu\***   **Xu Jiang\***   **Diogo Almeida\***   **Carroll L. Wainwright\***  
**Pamela Mishkin\***   **Chong Zhang**   **Sandhini Agarwal**   **Katarina Slama**   **Alex Ray**  
**John Schulman**   **Jacob Hilton**   **Fraser Kelton**   **Luke Miller**   **Maddie Simens**  
**Amanda Askell<sup>†</sup>**   **Peter Welinder**   **Paul Christiano<sup>\*†</sup>**  
**Jan Leike\***   **Ryan Lowe\***

OpenAI

# Human Feedback

- Human reward

SAN FRANCISCO,  
California (CNN) --  
A magnitude 4.2  
earthquake shook the  
San Francisco

...  
overturn unstable  
objects.

An earthquake hit  
San Francisco.  
There was minor  
property damage,  
but no injuries.

$$R(s_1) = 8.0$$

The Bay Area has  
good weather but is  
prone to  
earthquakes and  
wildfires.

$$R(s_2) = 1.2$$

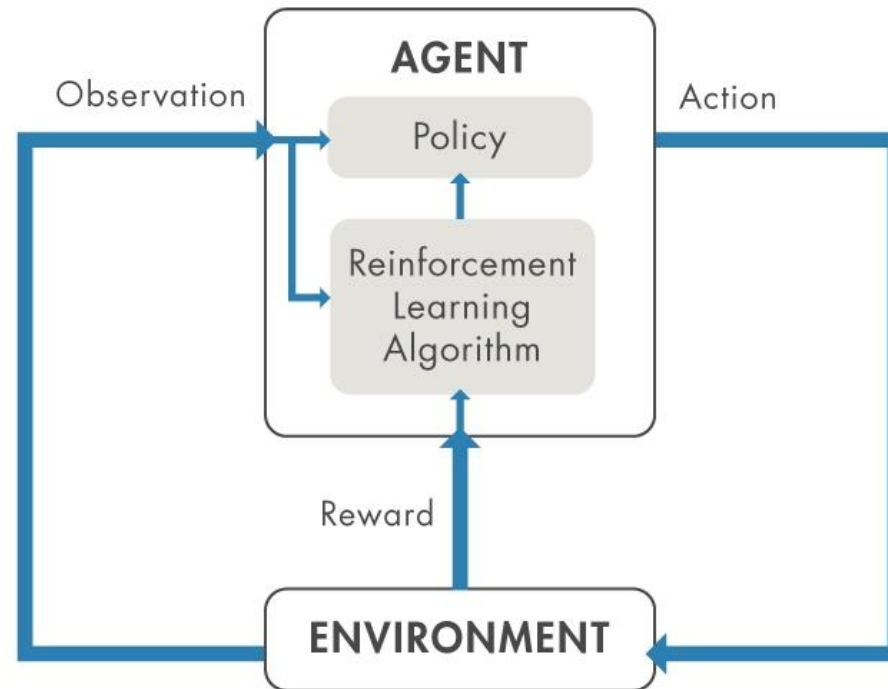
Goal: maximize the expected reward of samples from our LM

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})]$$

# Reinforcement Learning from Human Preferences

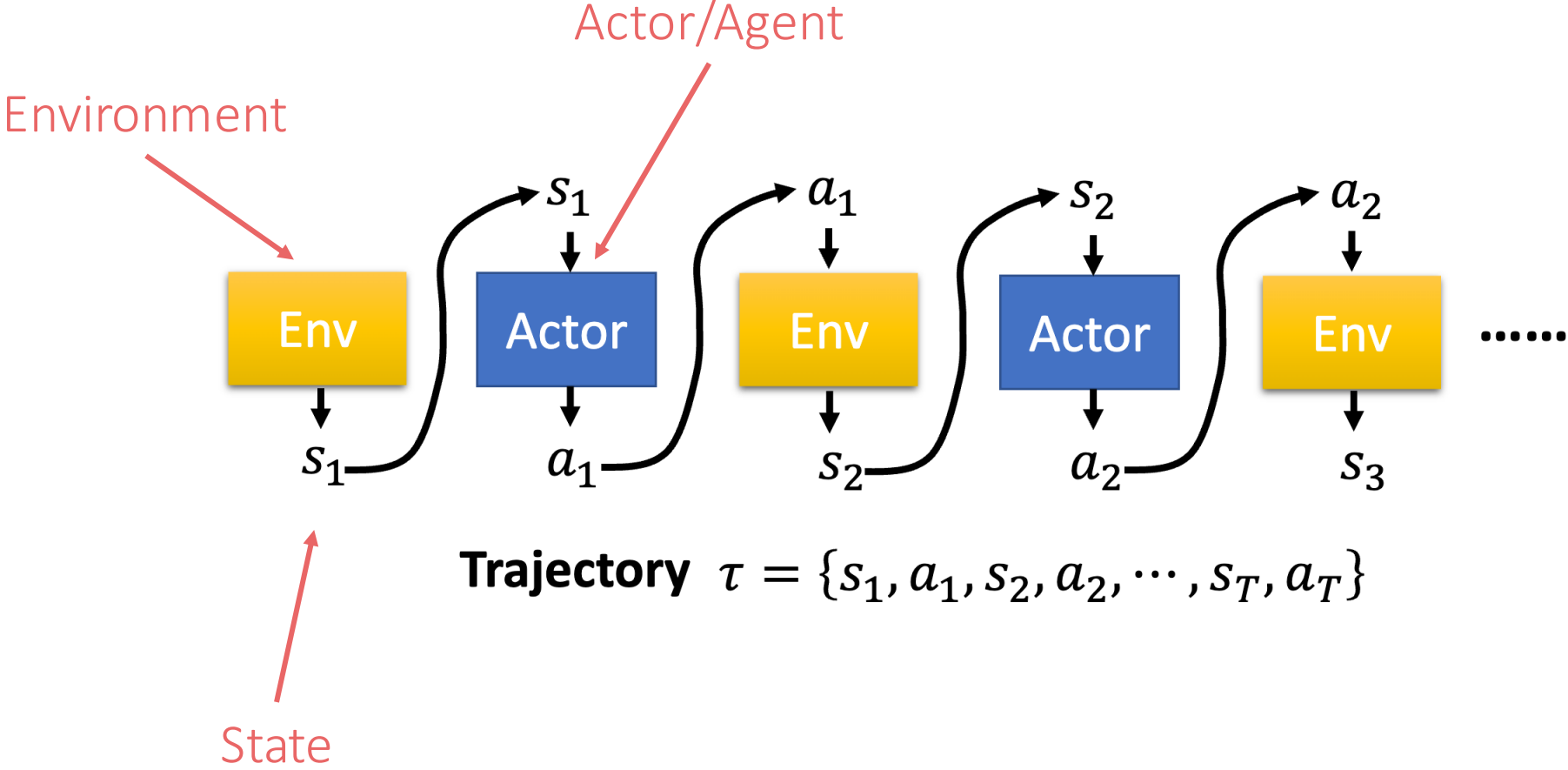
How do we change the LM parameters  $\theta$  to maximize this?

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})]$$

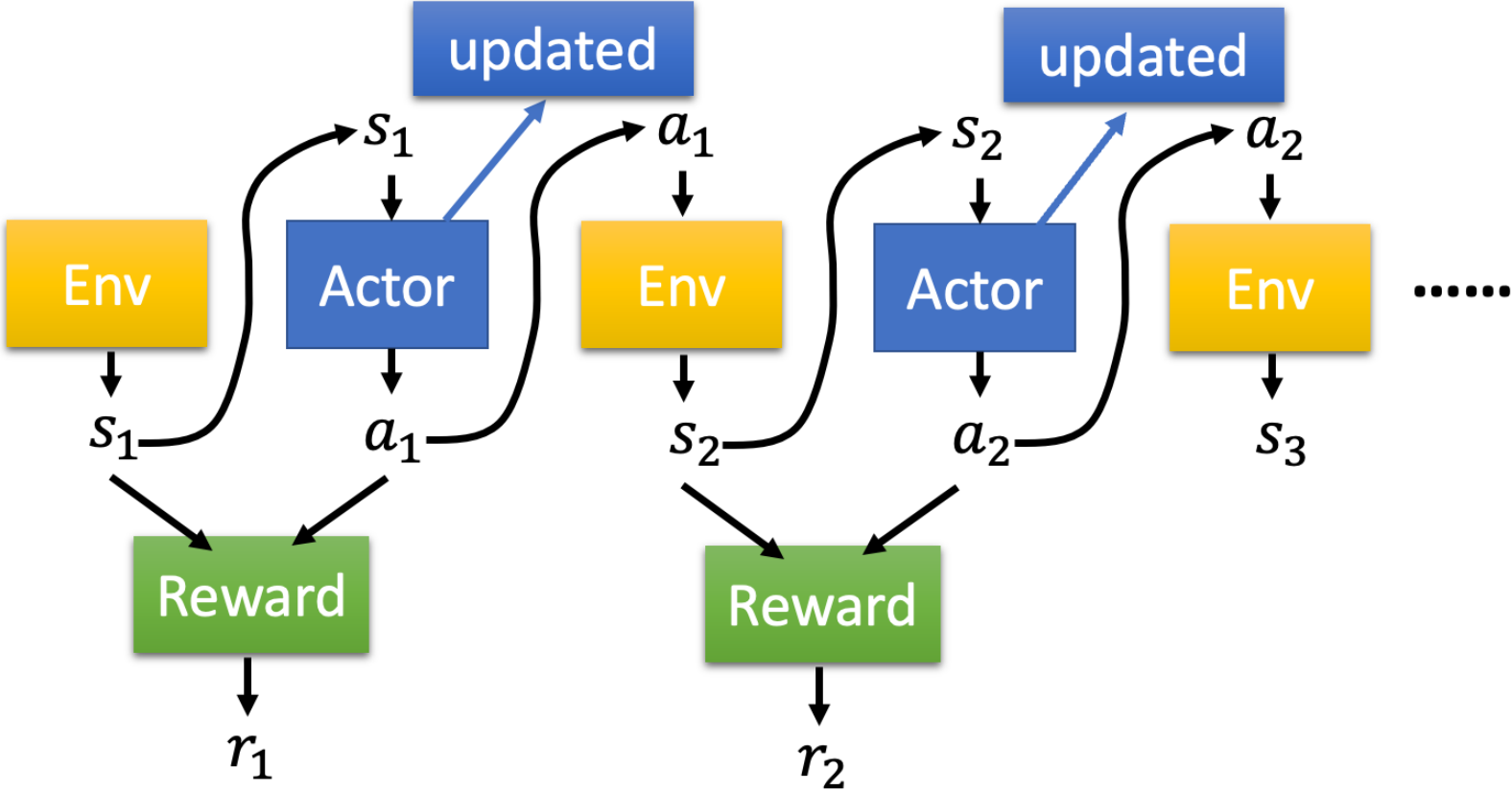




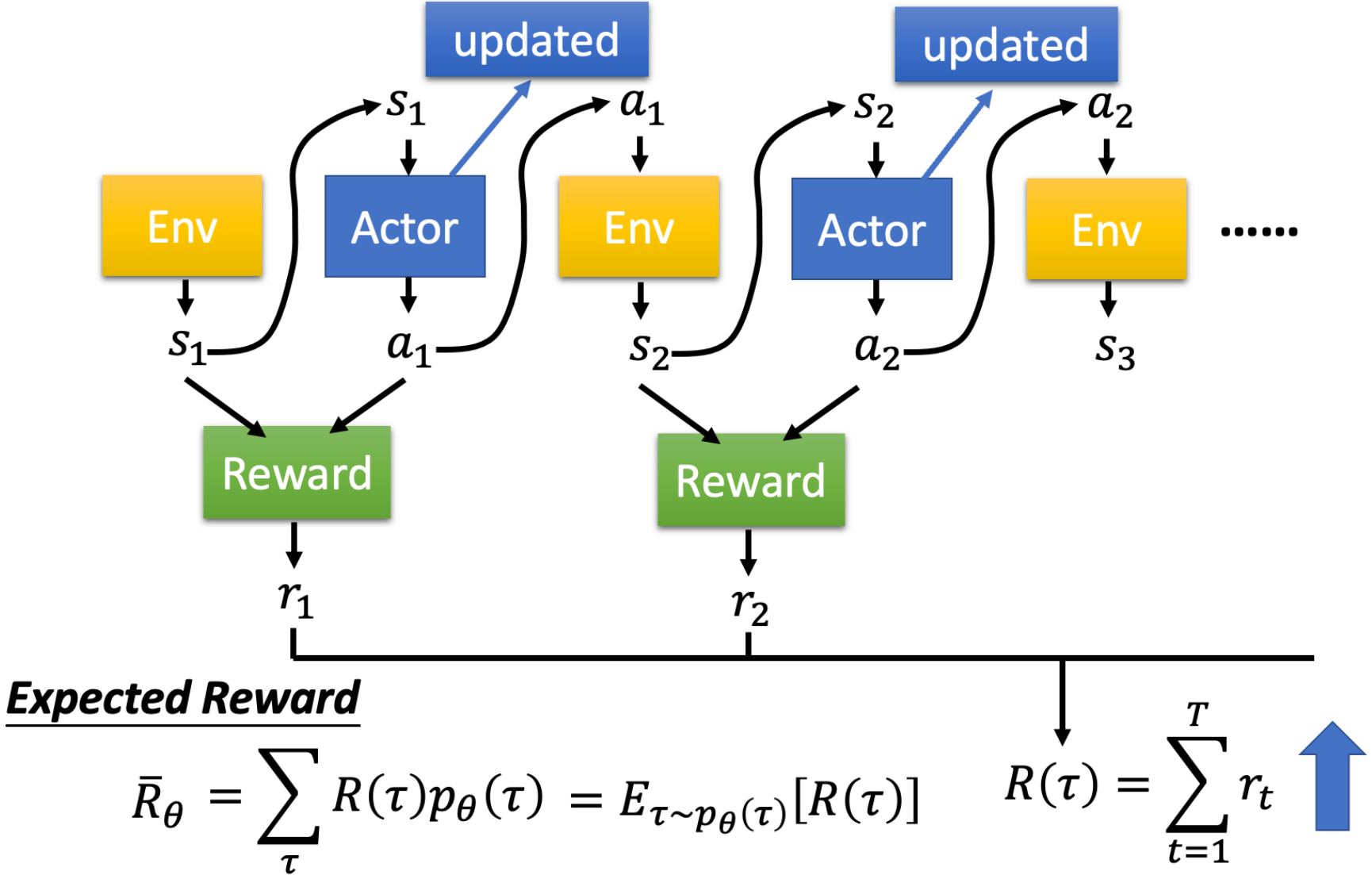
# Reinforcement Learning



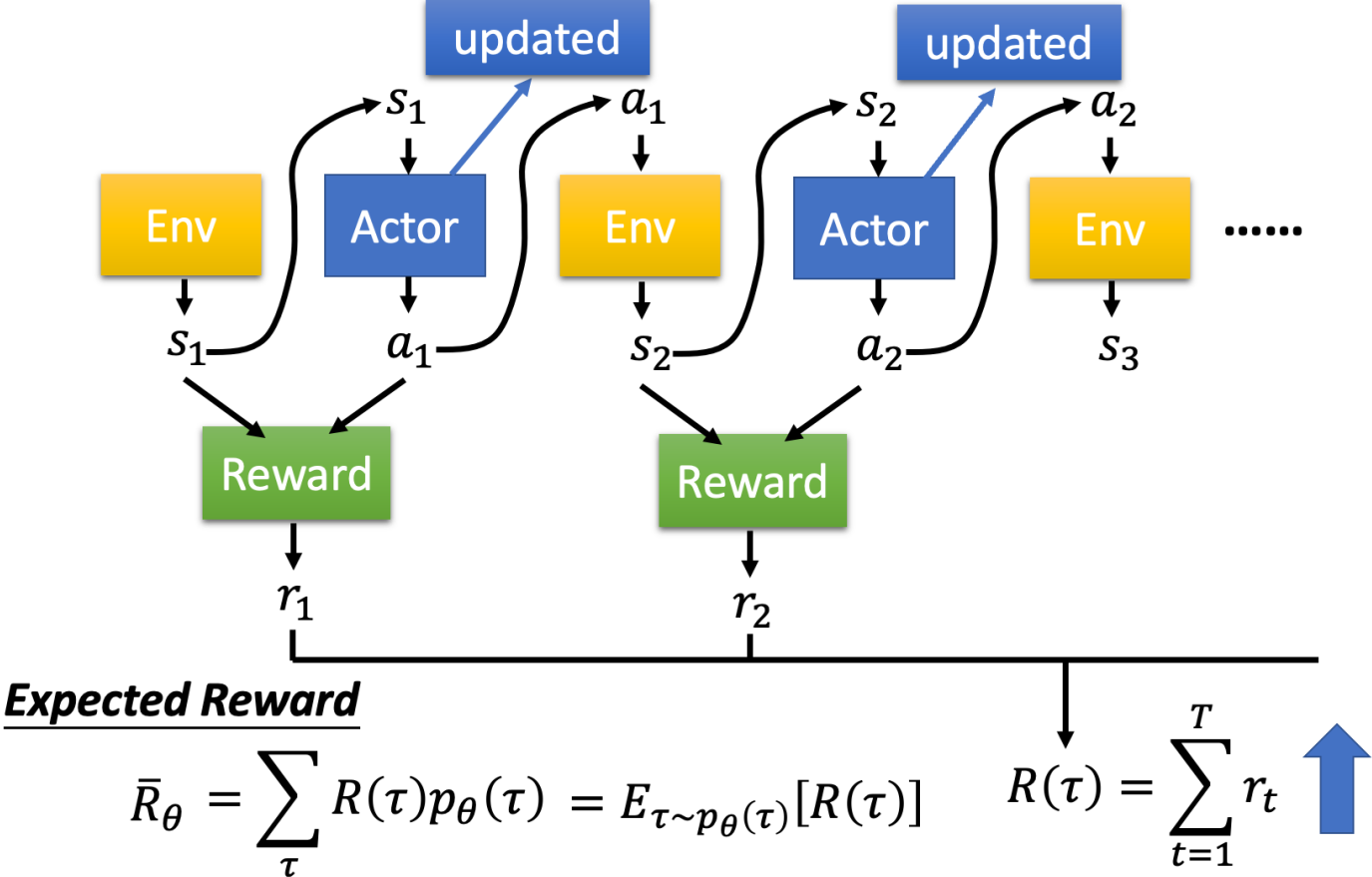
# Reinforcement Learning



# Reinforcement Learning



# Reinforcement Learning



- Solutions
- Q-Learning
  - Policy Gradient
  - Actor-Critic
  - ...

# Optimizing for Human Preferences

How do we change the LM parameters  $\theta$  to maximize this?

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})]$$

Gradient **Ascent**

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)} [R(\hat{s})]$$

**Policy Gradient** Methods in Reinforcement Learning  
(REINFORCE) [Williams, 1992]

# Policy Gradient/REINFORCE

Gradient Ascent

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)} [R(\hat{s})]$$

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \nabla_{\theta} \sum_s R(s) p_{\theta}(s) = \sum_s R(s) \nabla_{\theta} p_{\theta}(s)$$

Log-Derivative Trick

$$\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s) \quad \Rightarrow \quad \nabla_{\theta} p_{\theta}(s) = \nabla_{\theta} \log p_{\theta}(s) p_{\theta}(s)$$

# Policy Gradient/REINFORCE

$$\begin{aligned}\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] &= \sum_s R(s) \nabla_{\theta} p_{\theta}(s) = \sum_s p_{\theta}(s) R(s) \nabla_{\theta} \log p_{\theta}(s) \\ &= \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})]\end{aligned}$$

We can approximate this objective with Monte Carlo samples

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})] \approx \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$$

# Policy Gradient/REINFORCE

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

If  $R$  is +++

Take gradient steps to maximize  $p_{\theta}(s_i)$

If  $R$  is ---

Take steps to minimize  $p_{\theta}(s_i)$

We **reinforce** good actions, increasing the chance they happen again



# Proximal Policy Optimization (PPO)

- New parameters  $\theta'$  cannot be very different from old parameters  $\theta$

$$J_{PPO}^{\theta'}(\theta) = J^{\theta'}(\theta) - \beta K L(\theta, \theta')$$

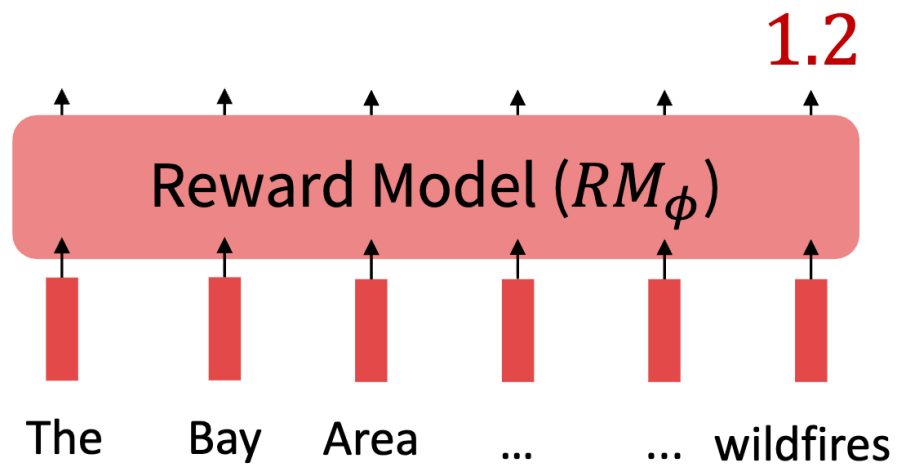
Regularization



- Initial policy parameters  $\theta^0$
- In each iteration
  - Using  $\theta^k$  to interact with the environment to collect  $\{s_t, a_t\}$  and compute advantage  $A^{\theta^k}(s_t, a_t)$
  - Find  $\theta$  optimizing  $J_{PPO}(\theta)$

# How to Model Human Preferences?

- Now for any reward function  $R$ , we can train our language model to maximize expected reward
- Problem 1: human-in-the-loop is expensive
  - Solution: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem
  - Train a reward model (RM) from an annotated dataset



# How to Model Human Preferences?

- Now for any reward function  $R$ , we can train our language model to maximize expected reward
- Problem 2: human judgments are noisy and miscalibrated
  - Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable

An earthquake hit San Francisco. There was minor property damage, but no injuries.

$S_1$

>

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

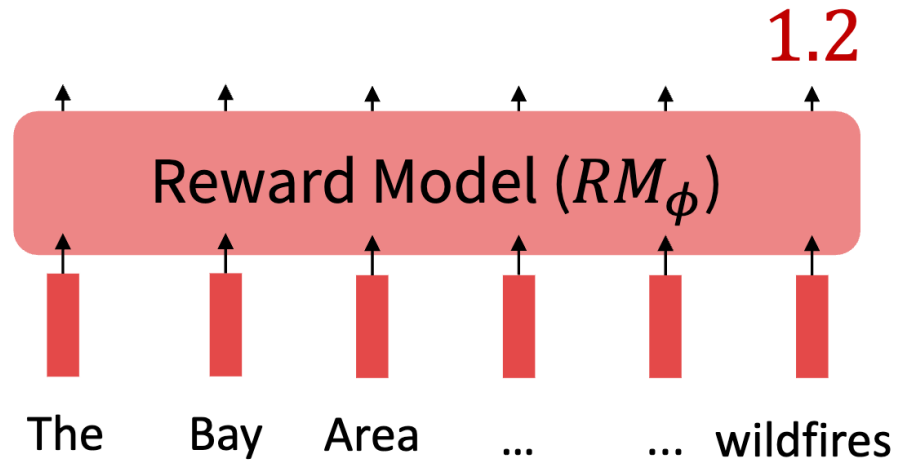
$S_3$

>

The Bay Area has good weather but is prone to earthquakes and wildfires.

$S_2$

# Training A Reward Model



Bradley-Terry [1952] paired comparison model

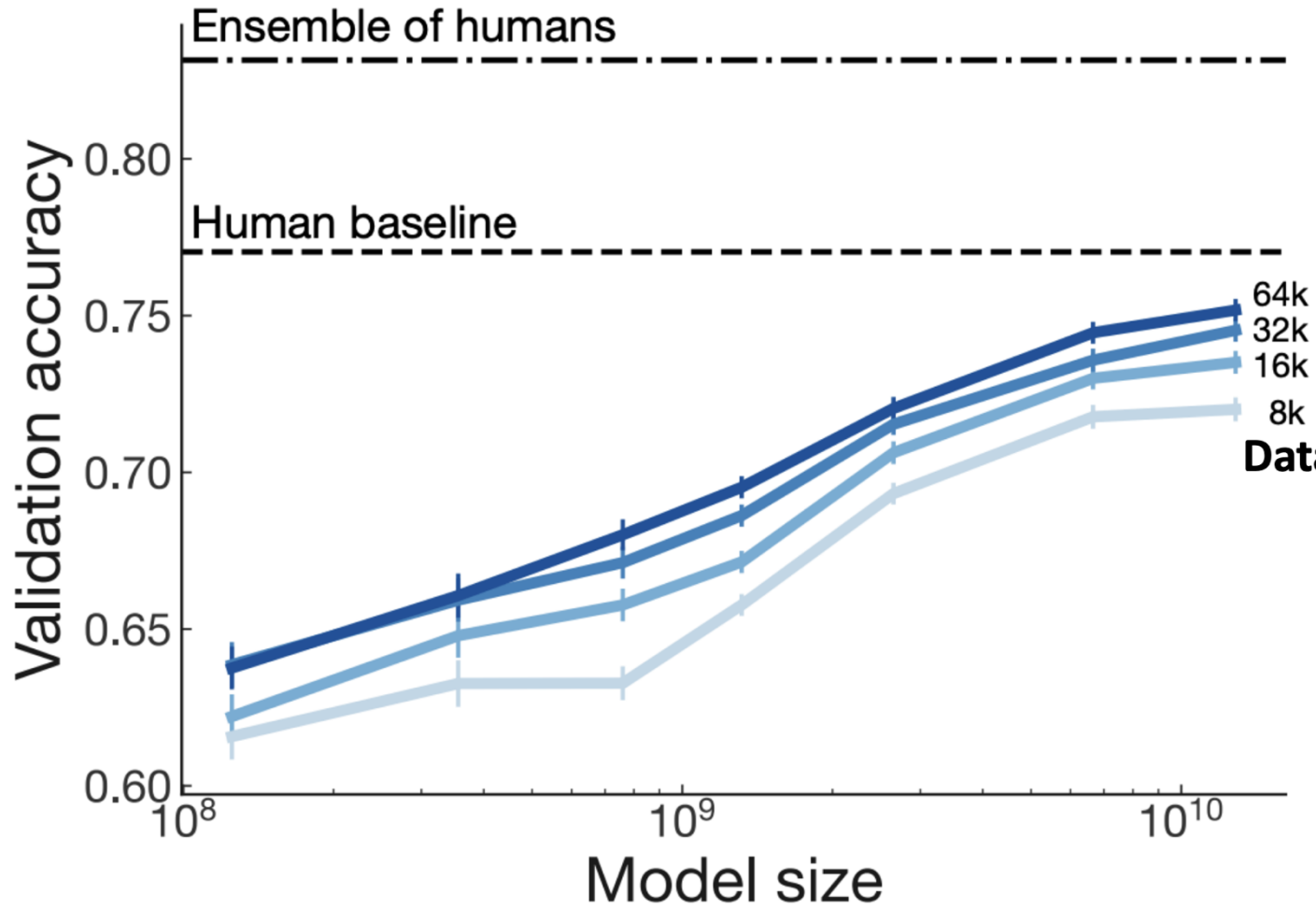
$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))]$$

“winning”  
sample

“losing”  
sample

$s^w$  should score  
higher than  $s^l$

# Reward Model vs. Real Human Feedback



Large enough RM trained on enough data approaching single human perf

[Stiennon et al., 2020]

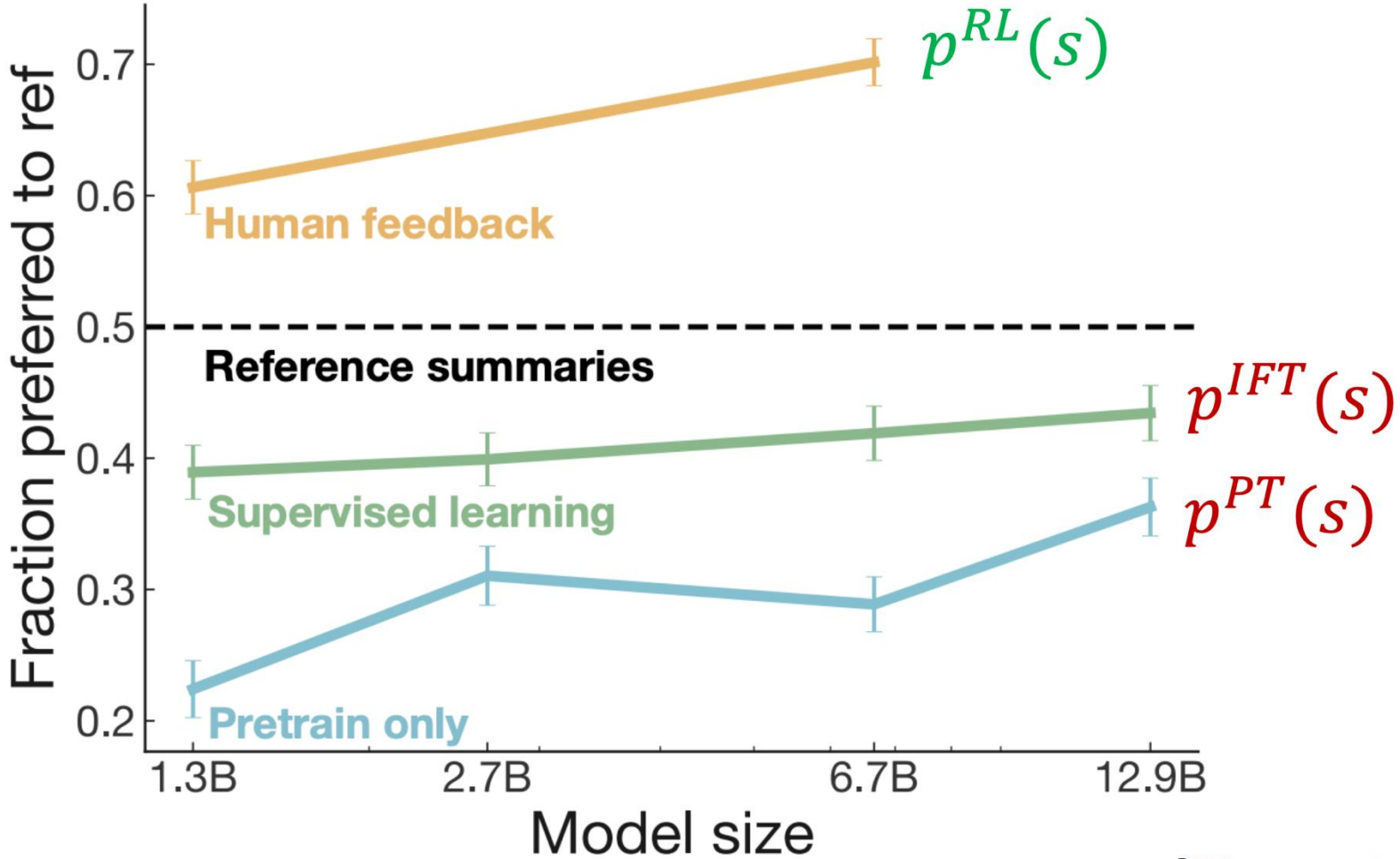
# RLHF: Putting Everything All Together

- Finally, we have everything we need:
  - A pretrained (possibly instruction-finetuned) LM  $p^{PT}(s)$
  - A reward model  $RM_\phi(s)$  that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
  - A method for optimizing LM parameters towards an arbitrary reward function.
- Now to do RLHF:
  - Initialize a copy of the model  $p_\theta^{RL}(s)$ , with parameters  $\theta$  we would like to optimize
  - Optimize the following reward with RL:

$$R(s) = RM_\phi(s) - \beta \log \left( \frac{p_\theta^{RL}(s)}{p^{PT}(s)} \right) \quad \text{Pay a price when } p_\theta^{RL}(s) > p^{PT}(s)$$

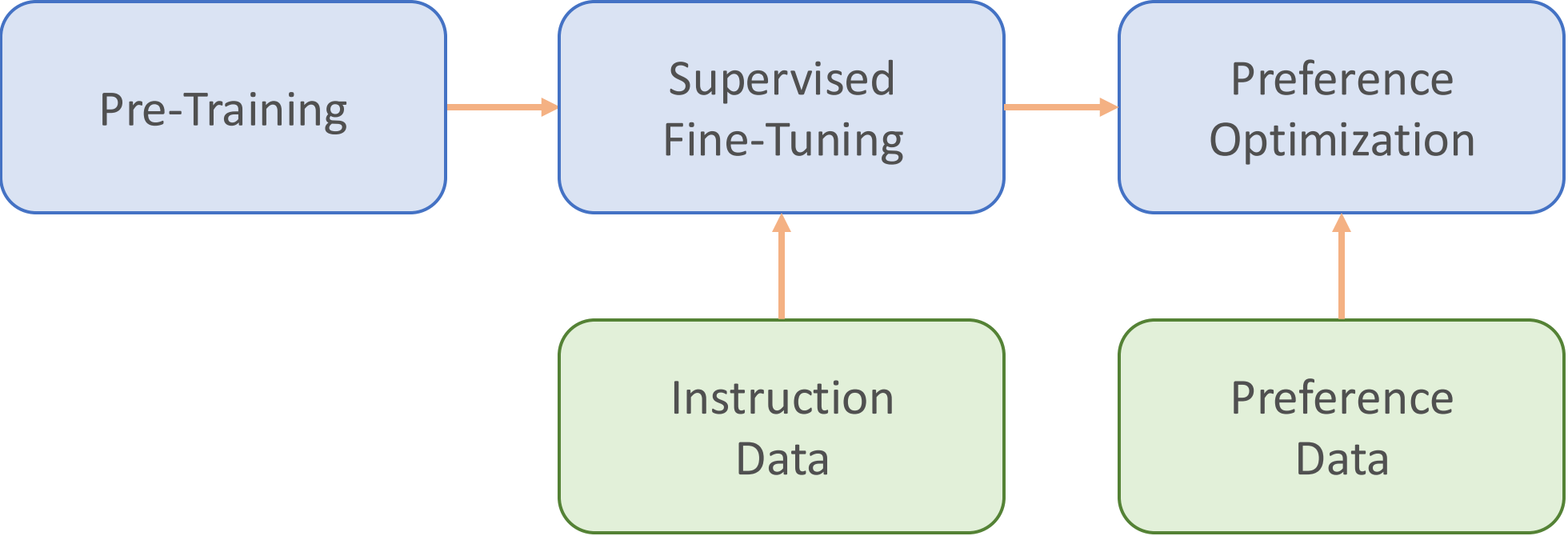
This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL) divergence** between  $p_\theta^{RL}(s)$  and  $p^{PT}(s)$ .

# RLHF vs. Supervised Fine-Tuning



[Stiennon et al., 2020]

# Alignment Pipeline



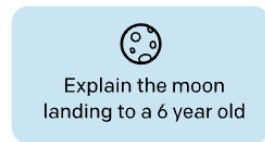


# InstructGPT

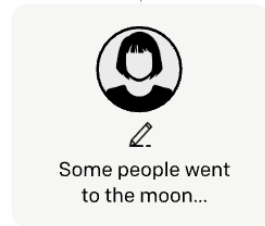
Step 1

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

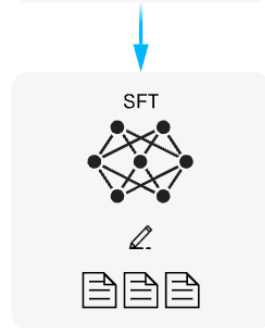
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



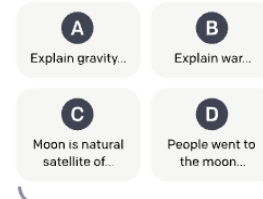
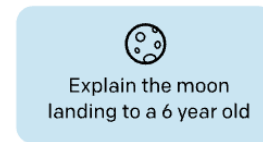
This data is used to fine-tune GPT-3 with supervised learning.



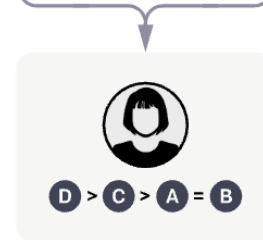
Step 2

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

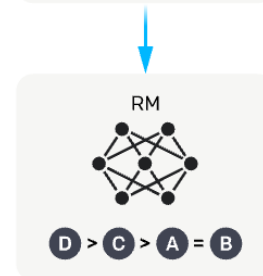
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



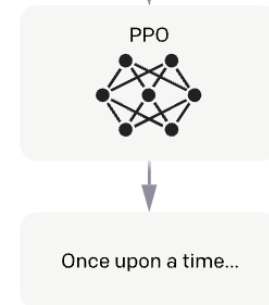
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

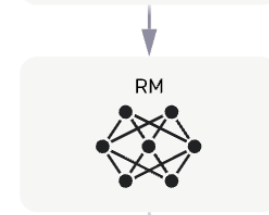
A new prompt is sampled from the dataset.



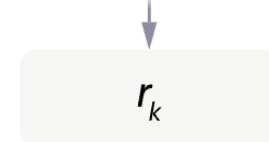
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



# InstructGPT

**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.

# ChatGPT: Instruction Fine-tuning + RLHF for Dialog Agents

## ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

## Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.