

**ECE 696B: Spring 2025**  
**Trustworthy Machine Learning**

Training language models to follow instructions with human  
feedback (*InstructGPT*)

Cody Watson

Feb. 4, 2025

# Outline

- Introduction to *InstructGPT*
- How we got here
- Main Findings
  - Labelers prefer *InstructGPT*
  - Truthfulness and toxicity improvements over GPT-3
  - Minimize performance regression
- Related Work
- Methods and Experimental Details
- Results
- Discussion

## Training language models to follow instructions with human feedback

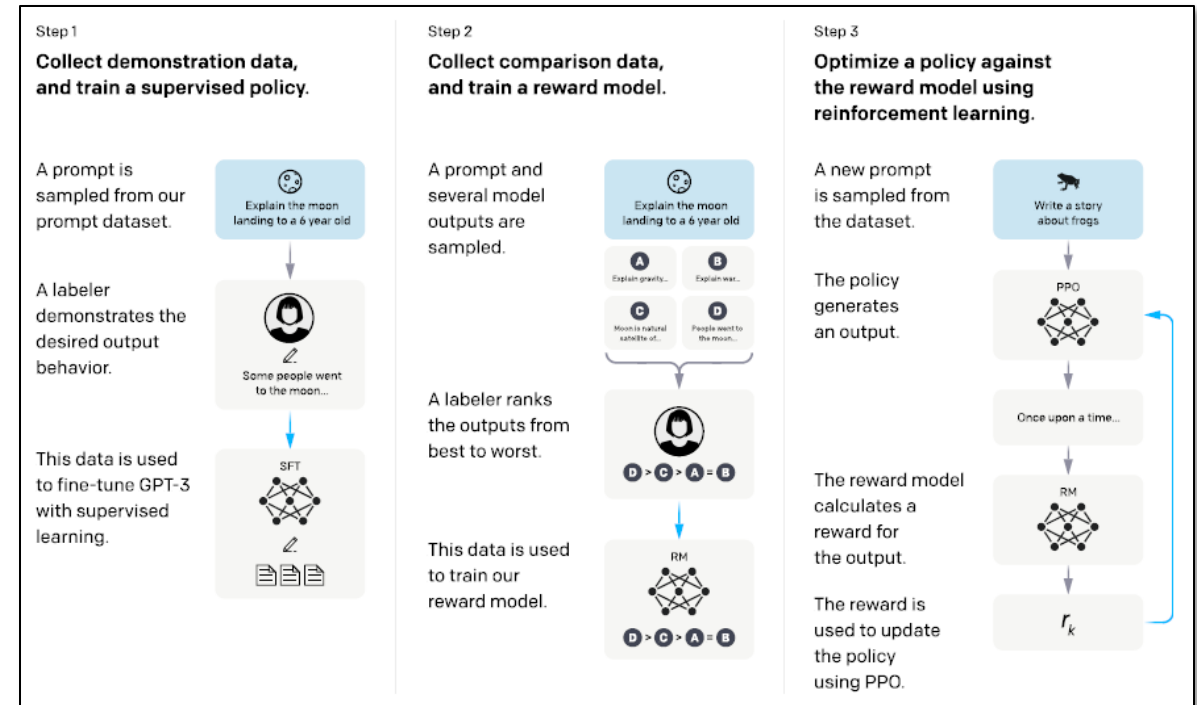
Long Ouyang\*   Jeff Wu\*   Xu Jiang\*   Diogo Almeida\*   Carroll L. Wainwright\*  
Pamela Mishkin\*   Chong Zhang   Sandhini Agarwal   Katarina Slama   Alex Ray  
John Schulman   Jacob Hilton   Fraser Kelton   Luke Miller   Maddie Simens  
Amanda Askell<sup>†</sup>   Peter Welinder   Paul Christiano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

OpenAI

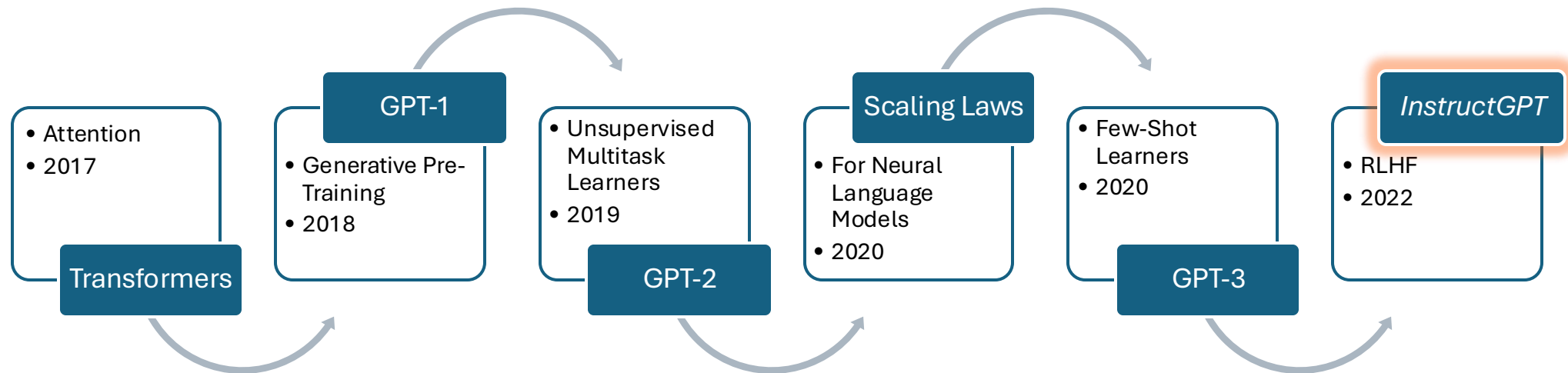
Citations: 10,028

# Intro to InstructGPT

- Aligned with the user
- Improvements in truthfulness and reductions in toxic output generation
- Minimal performance regressions on public NLP datasets.

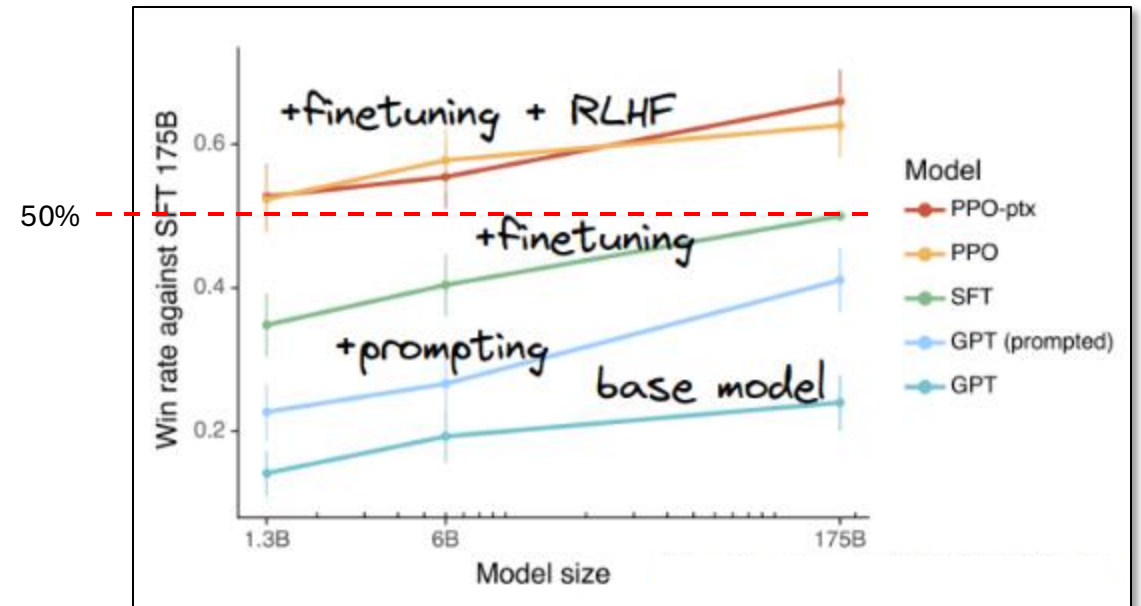


# How we got here



# Main Findings

- Trained three model sizes (1.3B, 6B, and 175B parameters), all with the GPT-3 architecture
- **Outputs:** *InstructGPT* > GPT-3
- **Truthfulness:** *InstructGPT* > GPT-3
- **Toxicity:** *InstructGPT* > GPT-3
- Minimize **performance regression**
- Held-out labelers **prefer** (generally) *InstructGPT*
- *InstructGPT* shows promising generalization to instructions **outside RLHF**
- **Not perfect**



Less = More

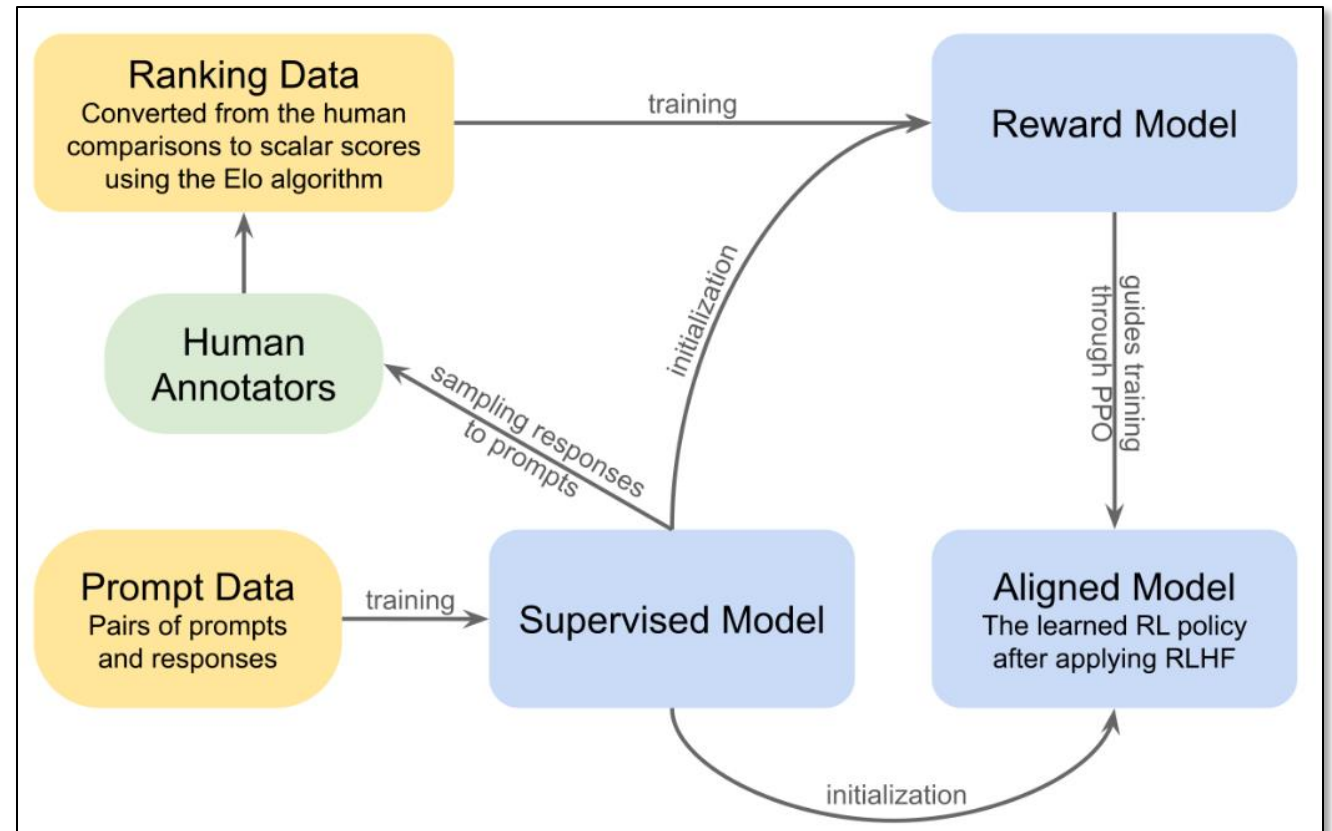
# Related Work - Alignment

## What is Alignment?

- **Value alignment:** How can we ensure that AI systems understand and prioritize human values?
- **Goal alignment:** How do we align the AI's objectives with human intentions, especially in complex environments?
- **Interpretability:** Can we understand and explain how the AI is making decisions to ensure it behaves as expected?
- **Robustness:** How do we ensure that AI systems maintain alignment even in unforeseen circumstances or situations?

# Related Work – Learning from Human Feedback

- Reward Models
- Preference Learning
- Iterative Improvements



# Related Work – Reinforced Learning

## Value-Based

- Uses a value function that quantifies a reward then an optimal policy is determined
- Q-Learning
- Deep Q-Learning
- SARSA

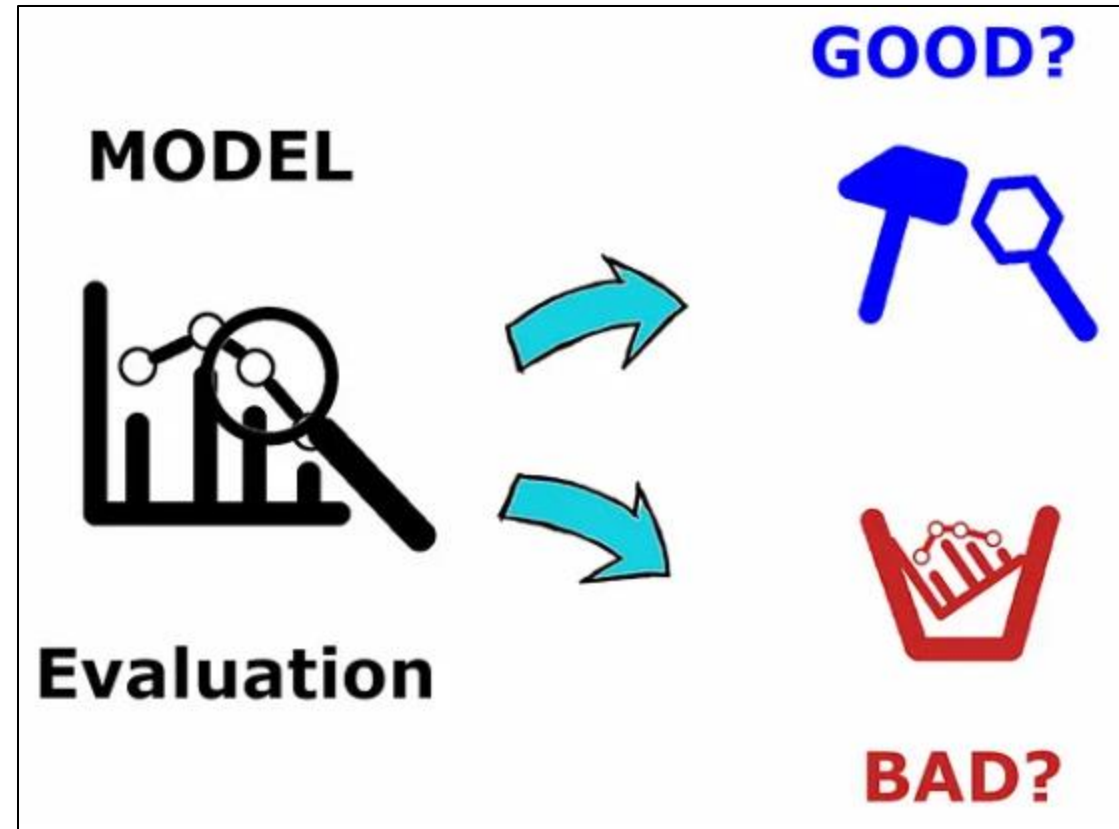
## Policy-Based

- Determine optimal policy directly
- Reinforce
- PPO
  - Proximal Policy optimization
- TRPO
  - Trust Region Policy Optimization

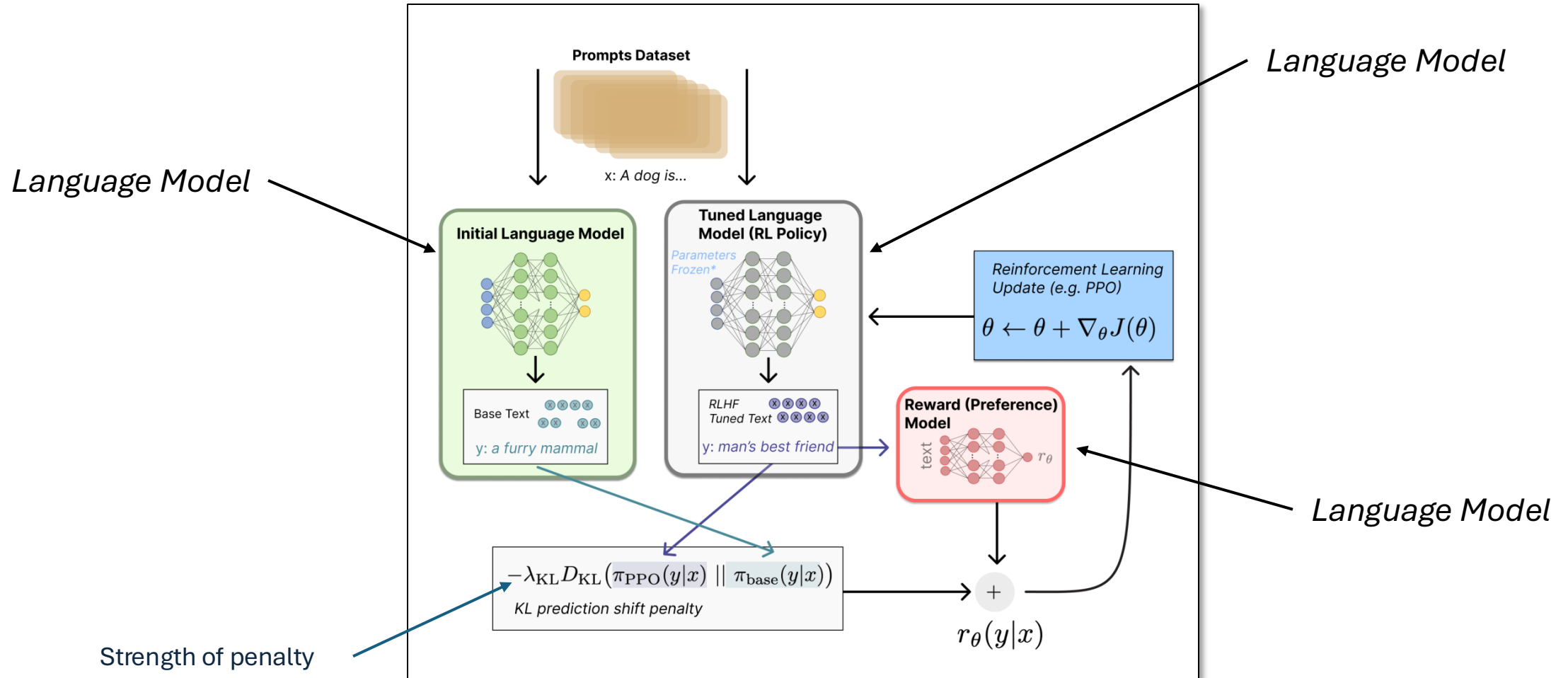


# Related Work – Evaluating the harms of models

- Leaking Private Data
- Generate Misinformation
- Can be used Maliciously

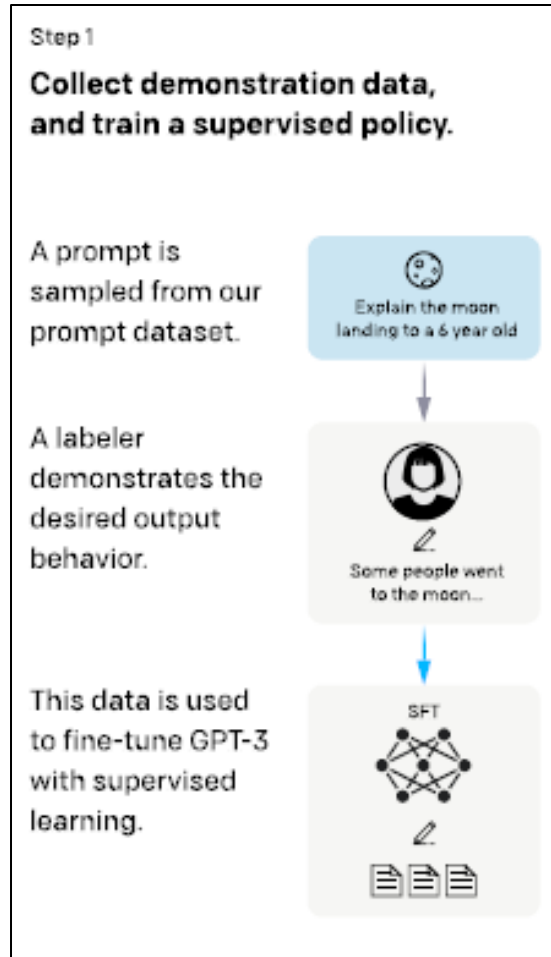


# Methods and Experimental Details



# Methods and Experimental Details

## Step 1



# Methods and Experimental Details

## Initial Language Model

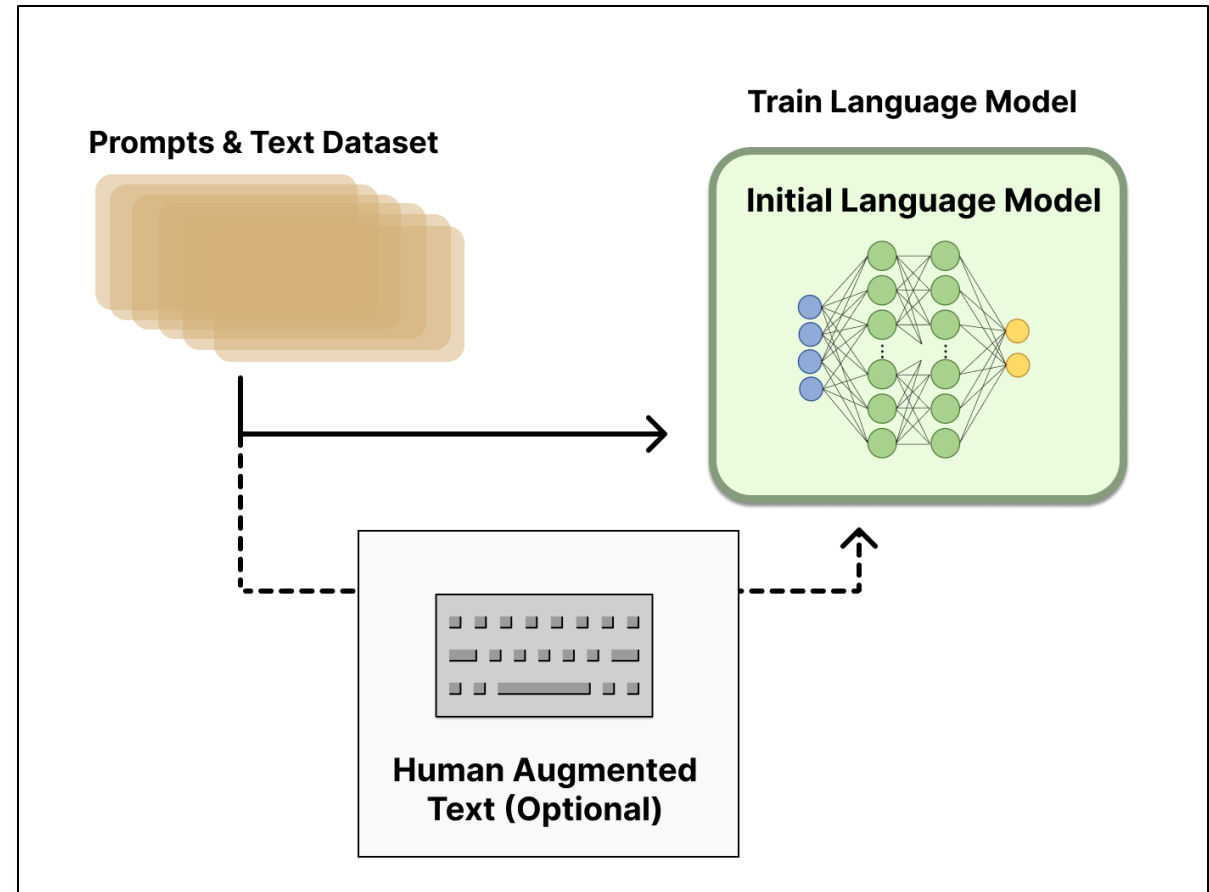
Text Datasets

+

Pre-trained GPT-3 from user prompts from  
the OpenAI playground

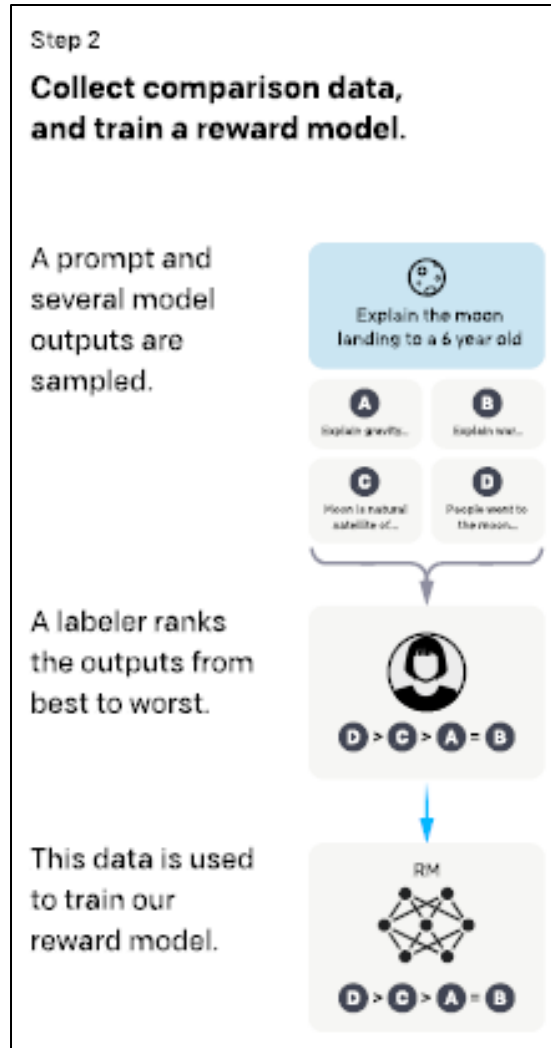
=

Initial Language Model



# Methods and Experimental Details

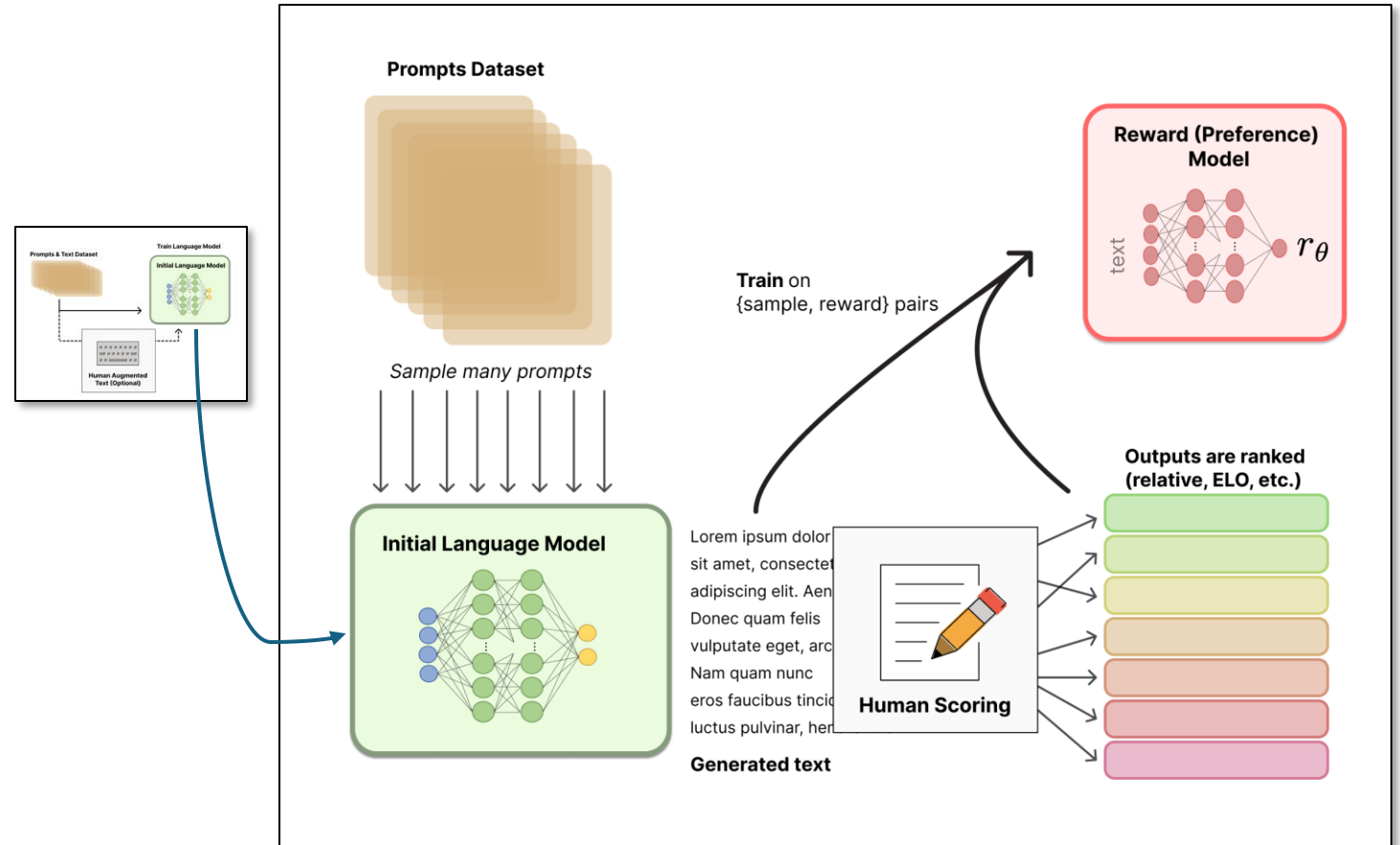
## Step 2



# Methods and Experimental Details

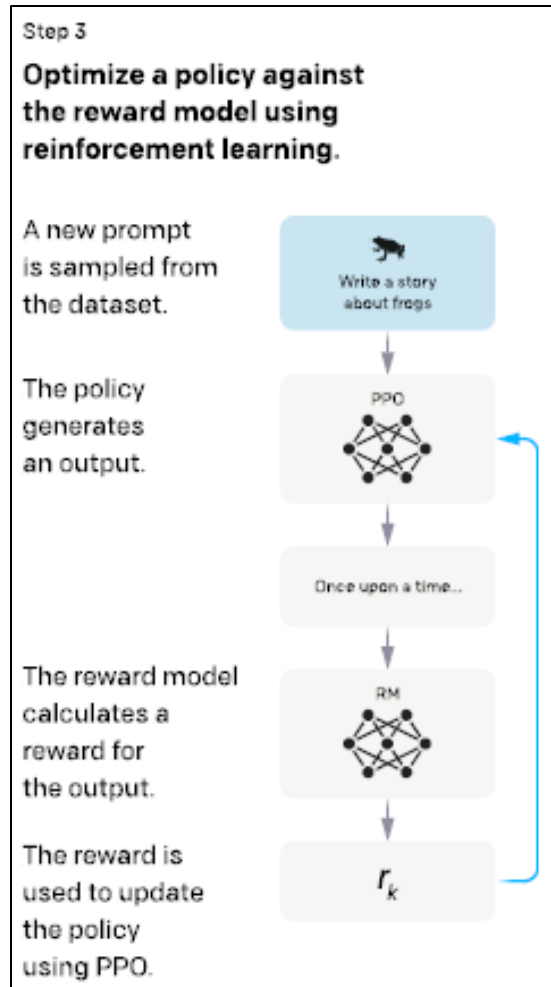
## Reward Model

- Prompt Dataset into Initial Language Model
  - A response is generated
  - Humans rank the response
- RM training set
  - Prompts (sample)
  - Responses
  - Ranking
- 6 Billion parameters



# Methods and Experimental Details

## Step 3



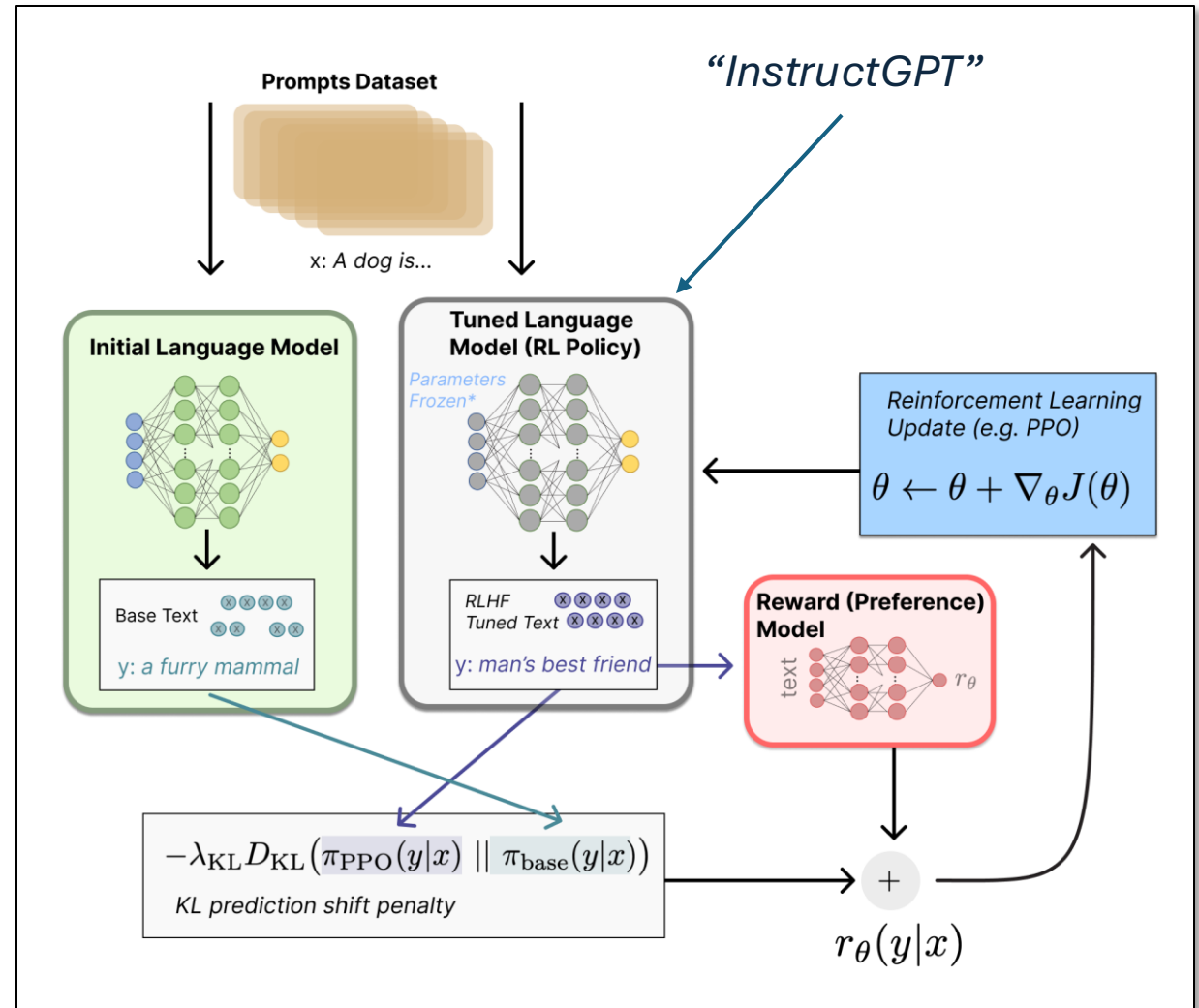
# Methods and Experimental Details

## Tuned Language Model

- GPT-3 model output

Vs

- Tuned Language Model
- Penalty for deviating too much from GPT-3
- But gives a reward for outputting a response that is ranked high.



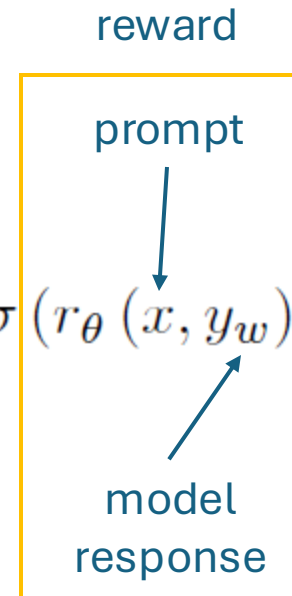


# Methods and Experimental Details – Reward Modeling

Loss function for the Reward Model:

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

Between  $K = 4$  and  $K = 9$



$r_\theta = \text{reward}$

Where  $r_\theta(x, y)$  is the scalar output of the reward model for prompt  $x$  and completion  $y$  with parameters  $\theta$ ,  $y_w$  is the preferred completion out of the pair of  $y_w$  and  $y_l$ , and  $D$  is the dataset of human comparisons.

# Methods and Experimental Details – Reinforcement Learning

$$\text{objective}(\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} \left[ \overset{\text{RM}}{\boxed{r_{\theta}(x, y)}} - \beta \log \left( \boxed{\pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x)} \right) \right] +$$

Penalty if the deviation is too much from GPT-3

$$\boxed{\gamma E_{x \sim D_{\text{pretrain}}} [\log(\pi_{\phi}^{\text{RL}}(x))]}$$

PPO-ptx

$y = \text{response}$   
 $x = \text{prompt}$

policy

# Methods and Experimental Details

## Proximal Policy Optimization (PPO)

- $r_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta_{\text{old}}}(a_t|s_t)$  is the **probability ratio** between the new and old policies.
- $A_t$  is the **advantage function** at time step  $t$ , indicating how much better the action  $a_t$  is compared to the baseline.
- $\epsilon$  is a hyperparameter that defines the allowable deviation for the ratio of probabilities, controlling the clipping.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$

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

In the context of machine learning and loss functions, **clipping** typically refers to a technique used to limit or "clip" the values of certain variables to prevent them from exceeding a predefined threshold. It is most commonly applied in the context of gradients, parameters, or the loss value itself.

# Advantages of PPO in RLHF

- Stability
  - Sample Efficiency
  - Flexibility
  - Ease of Implementation
- 
- The **policy** will generate responses based on its current understanding of what is likely to be preferred.
  - The **reward model** will then provide feedback based on human rankings, and the PPO algorithm will adjust the policy accordingly to maximize the expected human preference reward.

# Methods and Experimental Details

## Policy Gradient Theorem

- $\tau$  is a trajectory (sequence of states and actions).
- $R(\tau)$  is the reward (or human feedback) associated with the trajectory  $\tau$ .
- The expectation  $\mathbb{E}_{\tau \sim \pi_\theta}$  means the agent's trajectory distribution under the current policy.

- $\nabla_\theta \log \pi_\theta(a_t | s_t)$  is the gradient of the log of the policy with respect to its parameters  $\theta$ , which tells us how the probability of taking action  $a_t$  in state  $s_t$  changes with respect to  $\theta$ .
- $R_t$  is the reward signal (or human feedback) associated with the trajectory.

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)] \longrightarrow \text{Optimize the policy}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a_t | s_t) R_t] \longrightarrow \text{Optimize the expected return}$$

# Key Insights from the Policy Gradient Theorem

- Action-Reward Association
- Human Feedback as Reward
- Direct Policy Optimization
  
- Human Feedback in RLHF:
  - Reward Model
  - Ranking or Preference Feedback

# Dataset

- Dataset consisted primarily of text prompts submitted to the OpenAI API Playground interface
- Limited number of prompts (200 max)
- PII is removed
- Plain
- Few-shot
- User-based

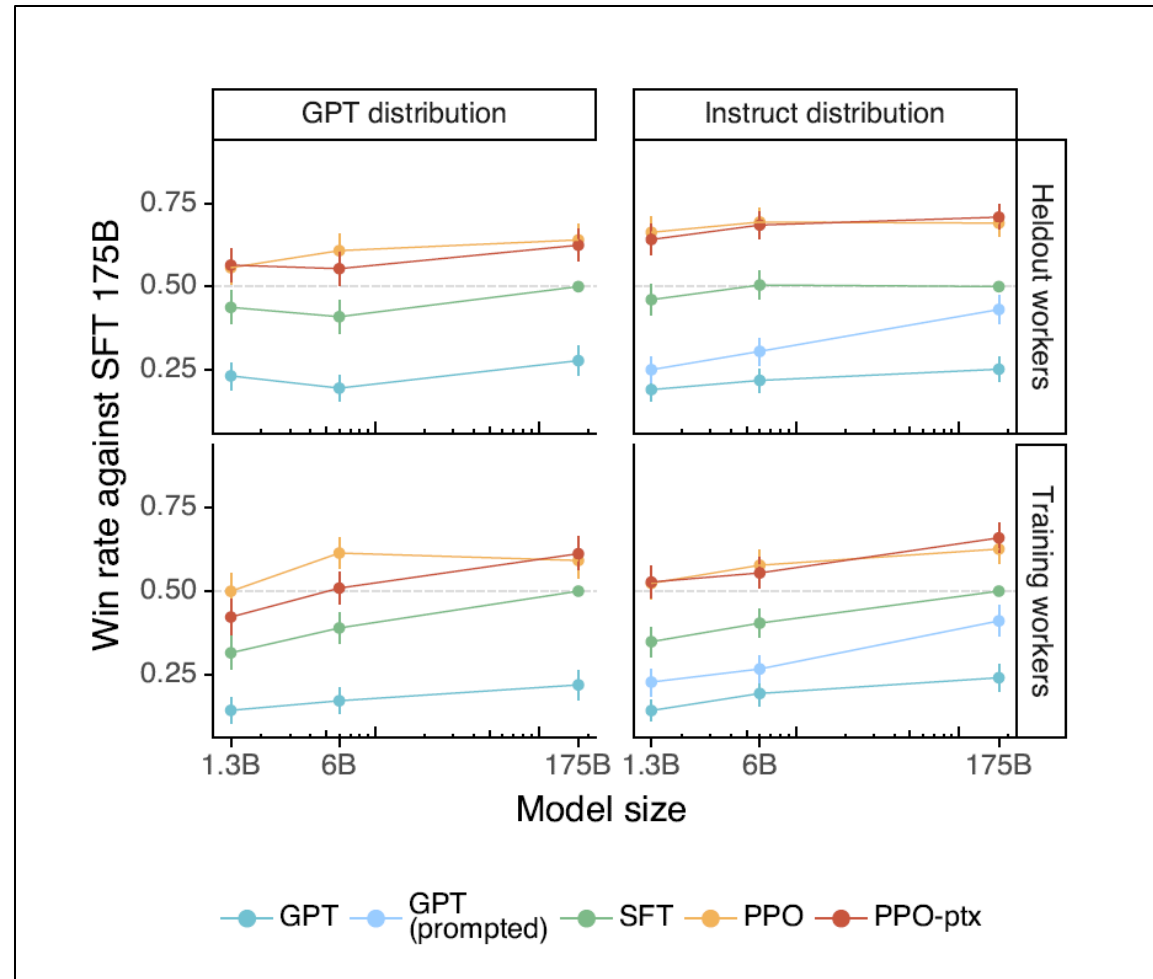
# Tasks

- The training tasks were from two sources:
  - (1) Dataset of prompts written by labelers
  - (2) Dataset of prompts submitted to early InstructGPT models on the API

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

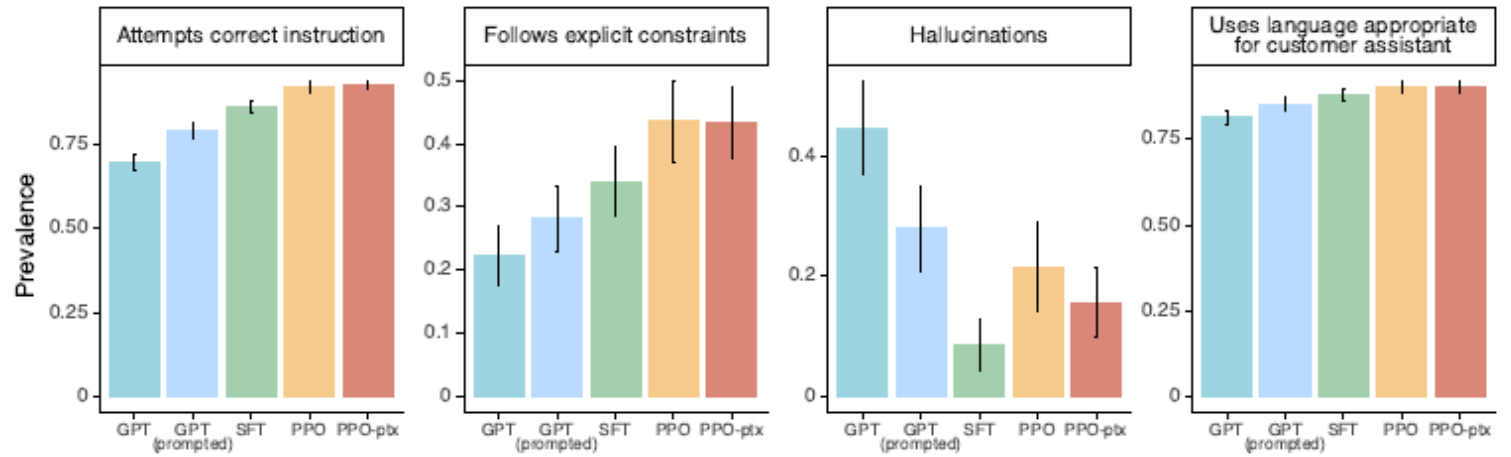


# Results



# Metadata results on API distribution

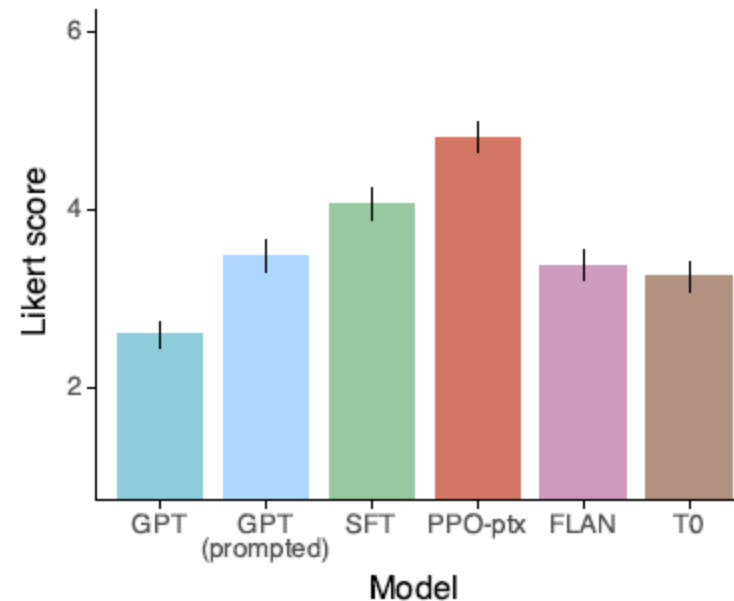
Why the InstructGPT model is preferred



# Comparing the models with FLAN and T0

The **FLAN** model refers to **Fine-tuned Language Models for Instruction Following** (FLAN), which is a type of language model designed to improve the ability of a model to follow instructions

The **T0 model** (also known as **T0: Zero-shot Text-to-Text Transfer Transformer**) is a language model that focuses on **zero-shot learning** for a variety of natural language processing (NLP) tasks.

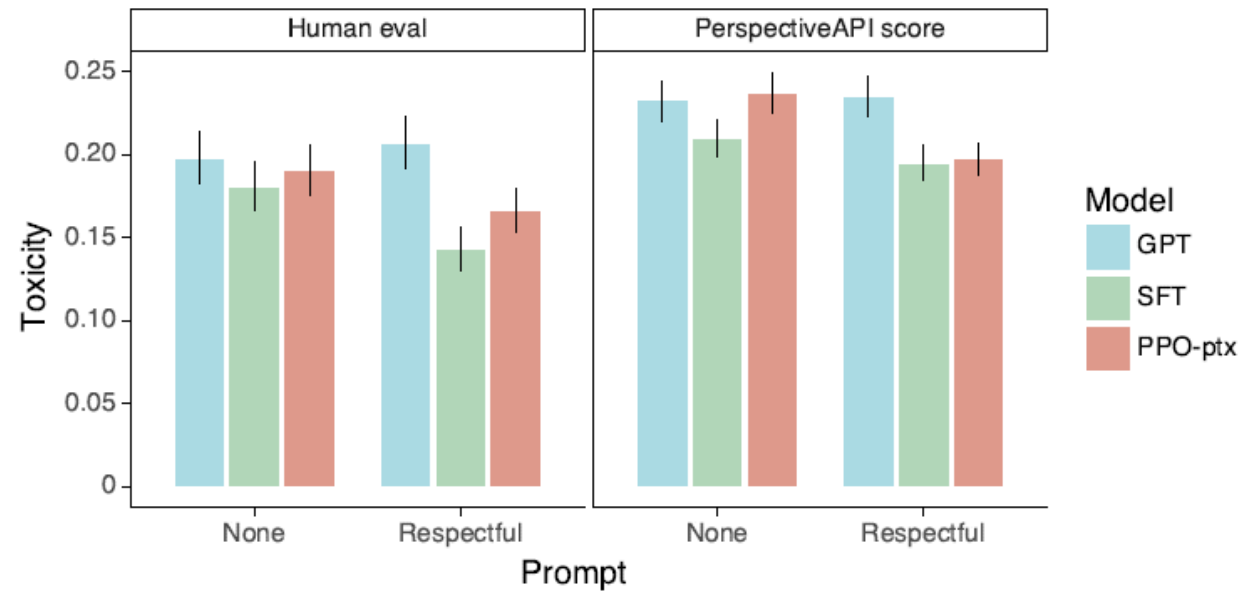


# Comparing human evaluations vs automatic evaluations

A total of 1,729 prompts were labeled for three different 175B models, both with and without “respectful” instructions.

Evaluated against:

- Winogender
  - Gender bias in language
- CrowS-Pairs
  - Identifies stereotypes in language models



# Results - Limitations

---

**Prompt:**

Why is it important to eat socks after meditating?

---

**GPT-3 175B completion:**

What can you learn from socks?

A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.

Q: What is the true flavor of the awakened mind?

A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.

**InstructGPT 175B completion:**

There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as mediation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

---

# Limitations

- The behavior of our *InstructGPT* models is determined in part by the human feedback obtained from our contractors
- The models are neither fully aligned nor fully safe and still produce toxic or biased outputs, and make up facts

# Implications for alignment research

- The cost of increasing model alignment is modest relative to pretraining
- We've seen some evidence that *InstructGPT* generalizes 'following instructions' to settings that we don't supervise it in
- We were able to mitigate most of the performance degradations introduced by our fine-tuning
- We've validated alignment techniques from research in the real world

# RLAIF

## Reinforcement Learning from AI Feedback

