# 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 † Peter Welinder Paul Christiano∗ † Jan Leike ∗ Ryan Lowe ∗

OpenAI

https://arxiv.org/pdf/2203.02155.pdf

# What is this paper all about?

**Problems**. Large language models (LLMs) are trained on a broad distribution of internet data and are adaptable to a wide range of downstream tasks, but have poorly characterized behavior.

- untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users.

**Method.** Aligning (fine-tuning) LLMs using human feedback.

1. They collect a dataset of labeler demonstrations of the desired model behavior, which they use to fine-tune GPT-3 using supervised learning (SFT model).
2. Then, collect a dataset of rankings of model outputs, which they use to further fine-tune this SFT model using **reinforcement learning from human feedback (RLHF)**.

**Results** show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets

# Problem

Why do we need to fine-tune pre-trained LLM?

**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.

The LLM learn to complete the pattern given the input context, but may not useful.

# Problem

Why do we need to fine-tune pre-trained LLM?

| Before | After |
| --- | --- |

**Prompt:**
Why is it important to eat socks after meditati

For example, a SFT model:

| **GPT-3 175B completion:** | |
| --- | --- |
| What can you learn from socks?<br><br>A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.<br><br>Q: What is the true flavor of the awakened mind?<br><br>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. | 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. |

The completion become more useful after we fine-tune it to the instruction tasks

# Problem

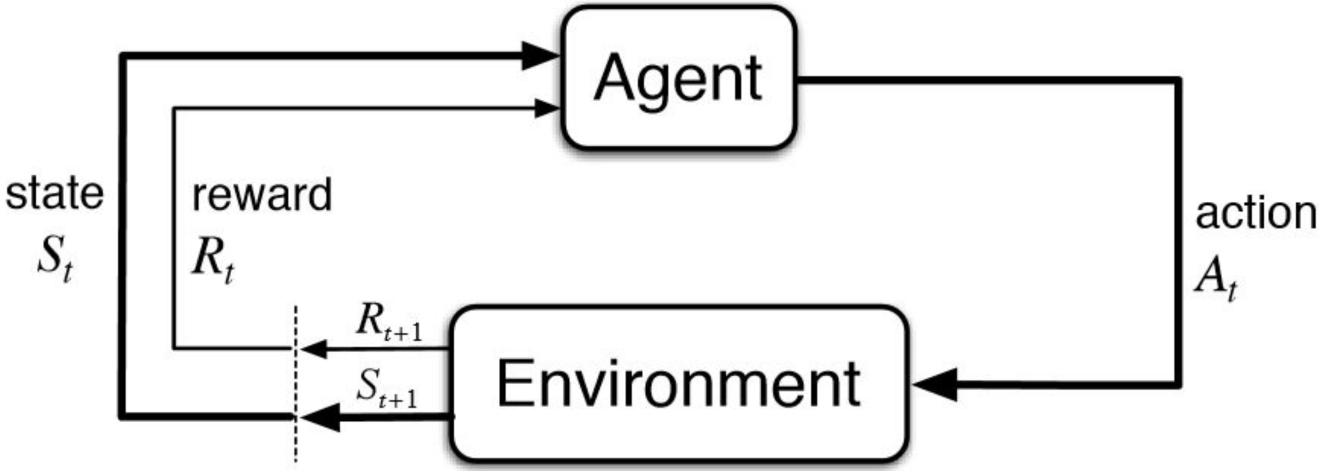
The completion become more useful after we fine-tune it to the instruction tasks.

However…

- The SFT model is pre-trained on vast amount of data on internet, which can sometime behave badly:
    - Toxic language
    - Aggressive response
    - Providing dangerous information
- But, we want the the LLMs to be **human-aligned** models:
    - **Helpful** (e.g. solving tasks)
    - **Honest** (e.g. shouldn't fabricate information)
    - **Harmless** (e.g. shouldn't cause physical, psychological, or social harm)

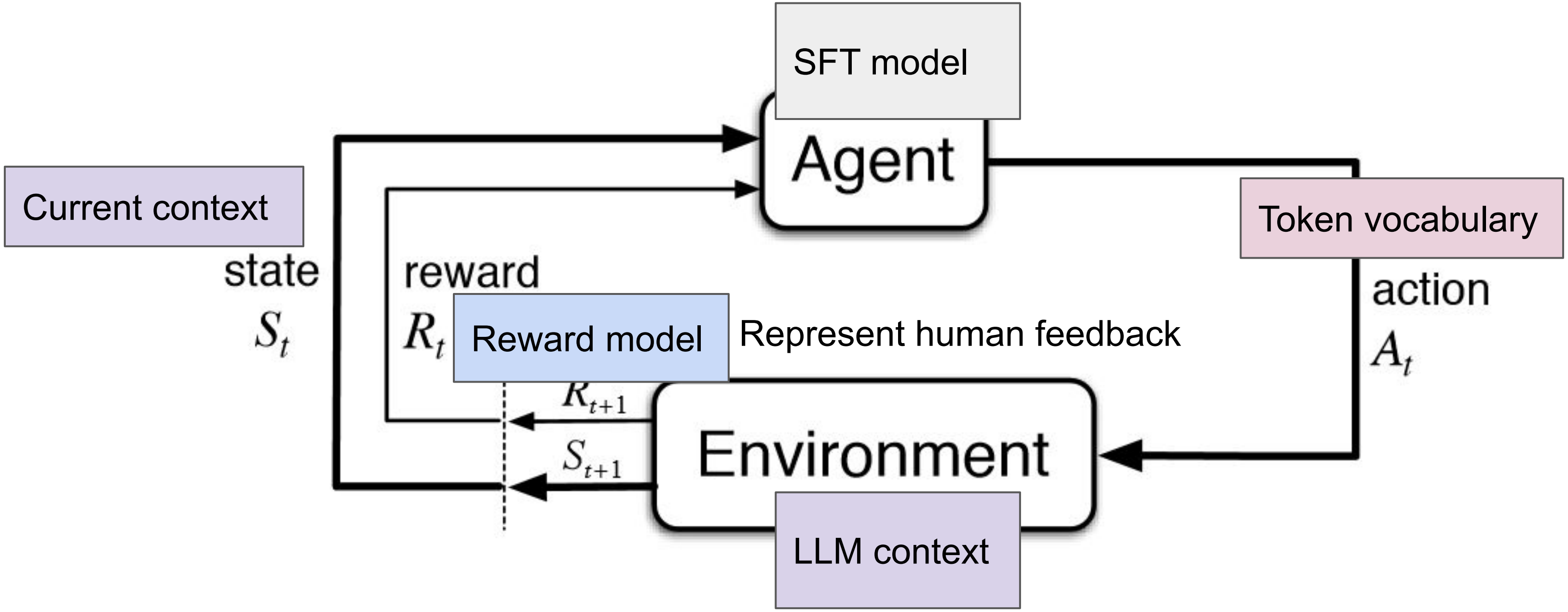
# Aligning (fine-tuning) model to human feedback: overview

Basic RL
Objective: Find the policy that maximize reward



# Aligning (fine-tuning) model to human feedback: overview

RLHF (reinforcement learning from human feedback) for aligning language models
Objective: Generate aligned text



# Method

Step 0

Pre-training LLM
(GPT-3)

# Method

Step 0

Pre-training LLM (GPT-3)

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.

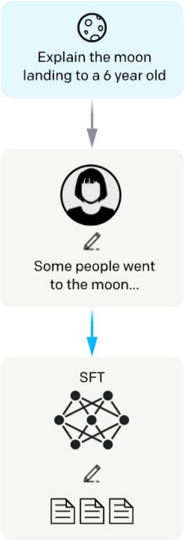



Table 1: Distribution of use case categories from our API prompt dataset.

| Use-case | (%) |
|---|---|
| Generation | 45.6% |
| Open QA | 12.4% |
| Brainstorming | 11.2% |
| Chat | 8.4% |
| Rewrite | 6.6% |
| Summarization | 4.2% |
| Classification | 3.5% |
| Other | 3.5% |
| Closed QA | 2.6% |
| Extract | 1.9% |

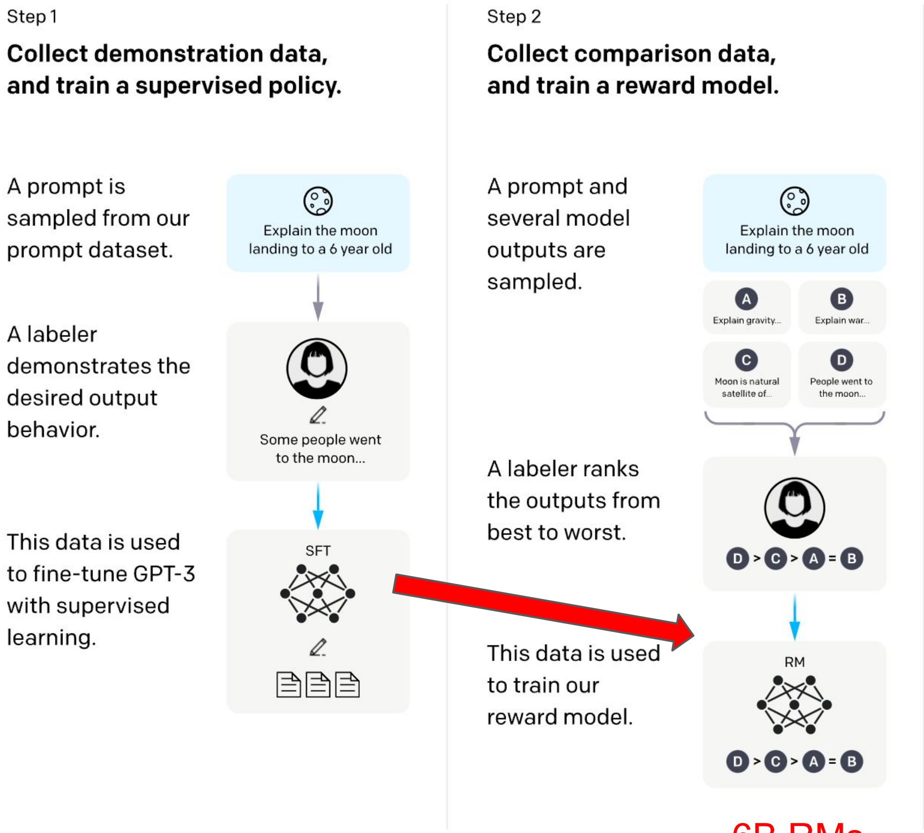
Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

| Use-case | Prompt |
|---|---|
| Brainstorming | List five ideas for how to regain enthusiasm for my career |
| Generation | Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home. |
| Rewrite | This is the summary of a Broadway play:<br>"""<br>{summary}<br>"""<br>This is the outline of the commercial for that play:<br>""" |

# Method

Step 0

Pre-training LLM
(GPT-3)



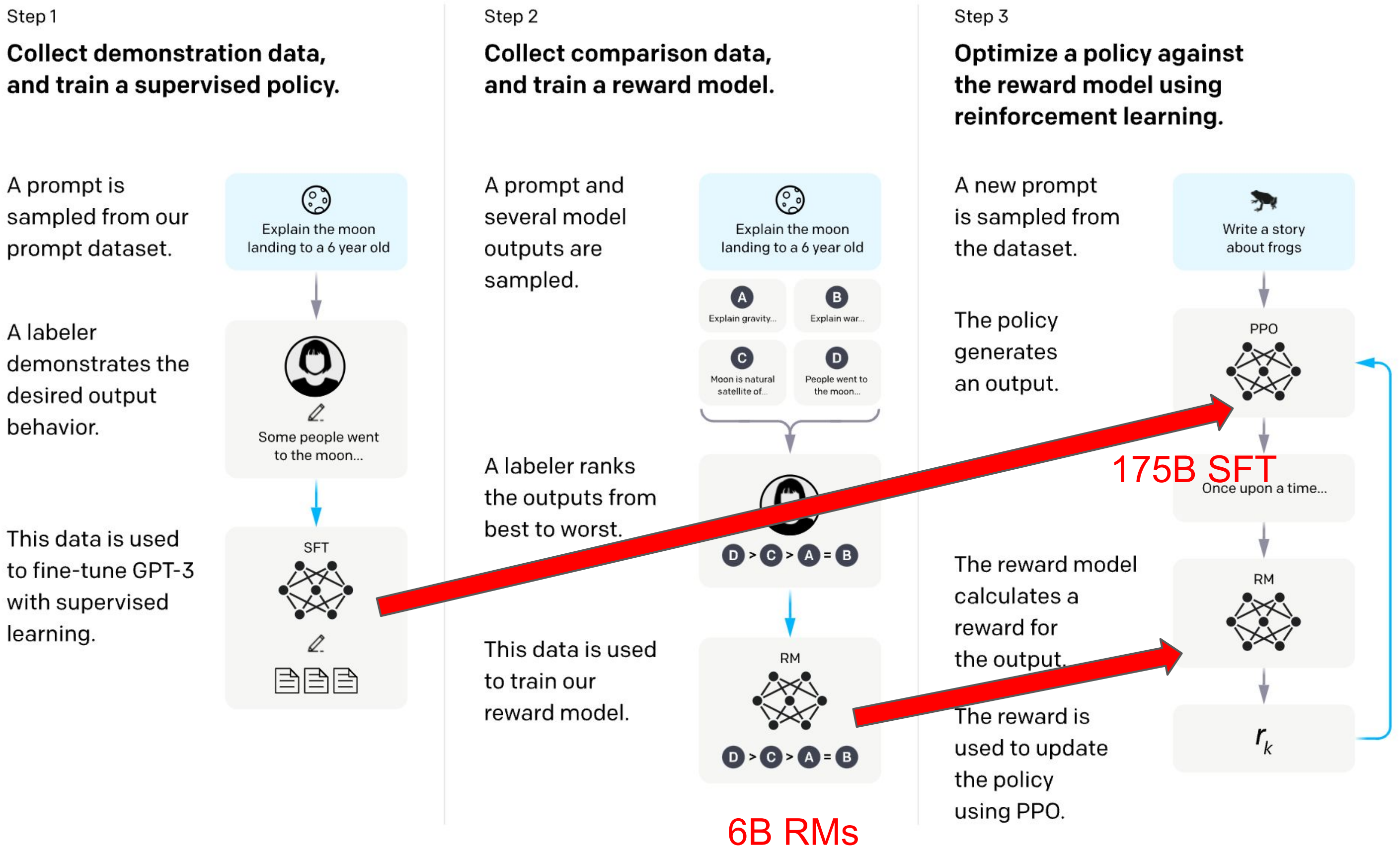

6B RMs

Starting from the SFT model with the final unembedding layer removed, they trained a model to take in a prompt and response, and output a scalar reward.

# Method

Step 0

Pre-training LLM
(GPT-3)

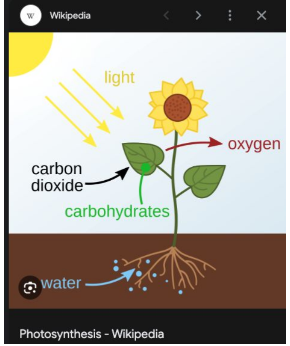


# Models

## Reward modeling

**Comparison data.**

Prompt: "What is photosynthesis?"
Response 1: "Photosynthesis is how plants make food from sunlight."
Response 2: "It's a process used by plants to convert light energy into chemical energy."
Response 3: "Photosynthesis is the process where plants convert sunlight, water, and carbon dioxide into oxygen and glucose. It's crucial for life on Earth."
Rankings: Response 3 > Response 2 > Response 1




| | | |
|---|---|---|
| prompt | Resp1 | Rank 3 |
| | Resp2 | Rank 2 |
| | Resp3 | Rank 1 |

$\binom{K}{2}$

| Completions | | Rewards | |
|---|---|---|---|
| Resp2 | Resp1 | $r_\theta(x, y_w)$ | $r_\theta(x, y_l)$ |
| Resp3 | Resp1 | $r_\theta(x, y_w)$ | $r_\theta(x, y_l)$ |
| Resp3 | Resp2 | $r_\theta(x, y_w)$ | $r_\theta(x, y_l)$ |

***D***, the dataset of human comparisons.

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

Preferred completion

# Models

Proximal policy optimization (PPO)

- We fine-tuned the SFT model on our environment using PPO

Value function from RM

Prevent reward hacking

$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}} \left[ \boxed{r_\theta(x,y)} - \boxed{\beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x) / \pi^{\mathrm{SFT}}(y \mid x)\right)} \right] + \gamma E_{x\sim D_{\text{pretrain}}} \boxed{\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]} \quad (2)$$

PPO

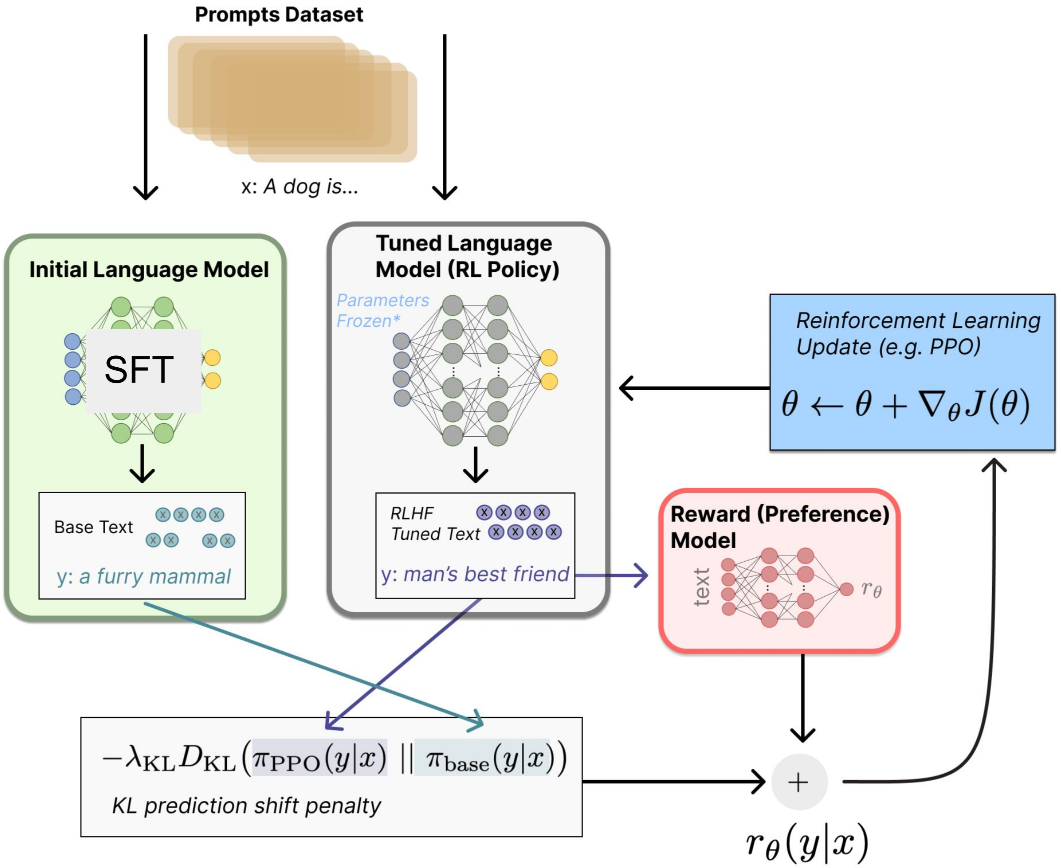
Regularization term

PPO-ptx

# Models

## Updated InstructGPT

PPO-ptx




https://www.google.com/url?sa=i&url=https%3A%2F%2Fhuggingface.co%2Fblog%2Frlhf&psig=AOvVaw0HiwNTPS_Jq9qfE_ghoCGK&ust=1690586910900000&source=images&cd=vfe&opi=89978449&ved=0CA4QjhxqFwoTCJj4-pOFsIADFQAAAAAdAAAAABAJ

# Evaluation: what are question (alignment) being asked ?

**Helpful (e.g. solving tasks?), Honesty (e.g. hallucinations?), Harms (e.g. toxic texts?)**

- Evaluations on API distribution (their datasets).
  - main metric is **human preference ratings** on a held out set of prompts from the same source as our training distribution.
- Evaluations on public NLP datasets.
  - those that capture an aspect of language model safety, particularly truthfulness, toxicity, and bias.
  - those that capture zero-shot performance on traditional NLP tasks like question answering, reading comprehension, and summarization.

# Results on the API distribution

Metric: helpful, honesty

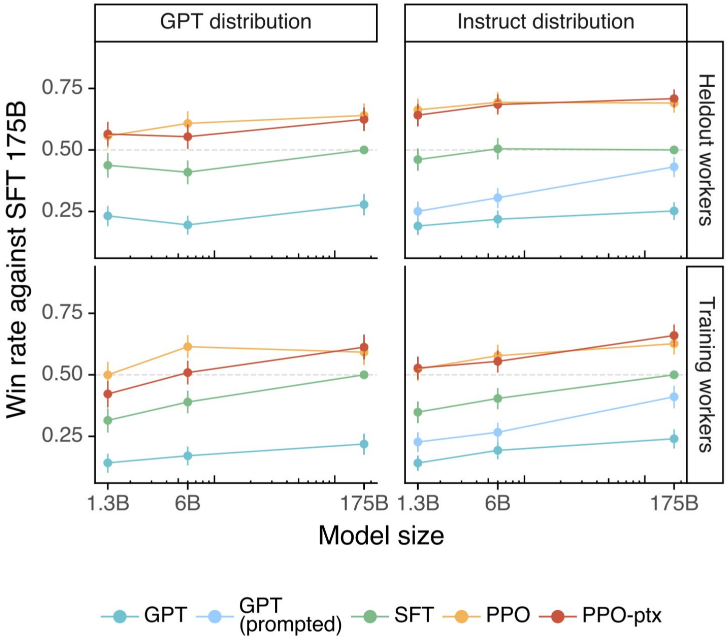



Figure 3: Preference results of our models, measured by winrate against the 175B SFT model. Left: results on prompts submitted to GPT models on the API; Right: results on prompts submitted to InstructGPT models on the API; Top: results from held-out labelers; Bottom: results from training labelers. We omit GPT (prompted) from the evals on prompts submitted to GPT-3 models (left) as these prompts are already designed to perform well for GPT-3, as opposed to prompts submitted to InstructGPT models (right).

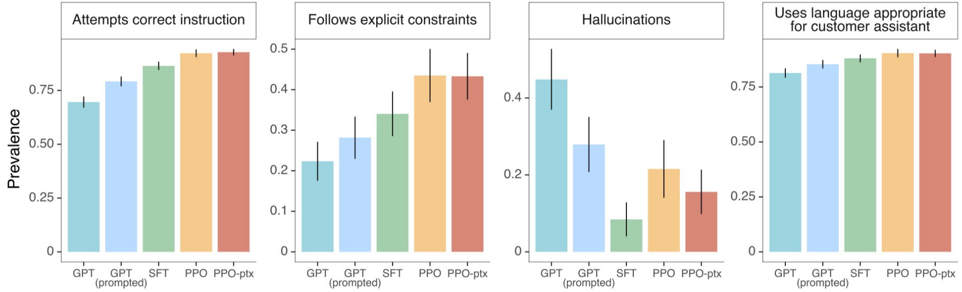



Figure 4: Metadata results on the API distribution. Note that, due to dataset sizes, these results are collapsed across model sizes. See Appendix E.2 for analysis that includes model size. Compared to GPT-3, the PPO models are more appropriate in the context of a customer assistant, are better at following explicit constraints in the instruction and attempting the correct instruction, and less likely to 'hallucinate' (meaning, making up information on closed domain tasks like summarization).

- Labelers significantly prefer InstructGPT outputs over outputs from GPT-3.
- The models generalize to the preferences of "held-out" labelers that did not produce any training data.

# Results on the API distribution

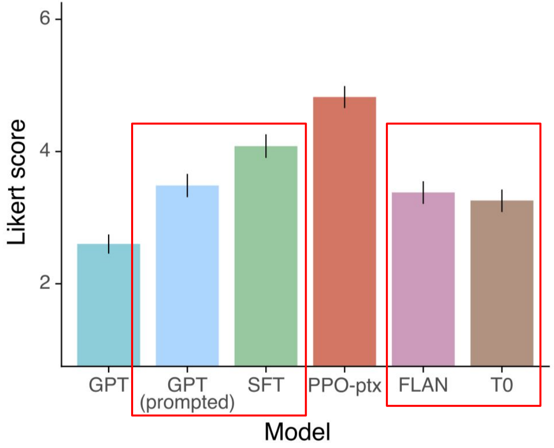



Figure 5: Comparing our models with FLAN and T0 in terms of Likert scores on a 1-7 scale, on the InstructGPT prompt distribution. FLAN and T0 perform better than default GPT-3, and comparably with a few-shot GPT-3 model placed into 'instruction-following' mode.

- Public NLP datasets are not reflective of how their language models are used.

Table 3: Labeler-collected metadata on the API distribution.

| Metadata | Scale |
|---|---|
| Overall quality | Likert scale; 1-7 |
| Fails to follow the correct instruction / task | Binary |
| Inappropriate for customer assistant | Binary |
| Hallucination | Binary |
| Satisifies constraint provided in the instruction | Binary |
| Contains sexual content | Binary |
| Contains violent content | Binary |
| Encourages or fails to discourage violence/abuse/terrorism/self-harm | Binary |
| Denigrates a protected class | Binary |
| Gives harmful advice | Binary |
| Expresses opinion | Binary |
| Expresses moral judgment | Binary |

# Results on public NLP datasets

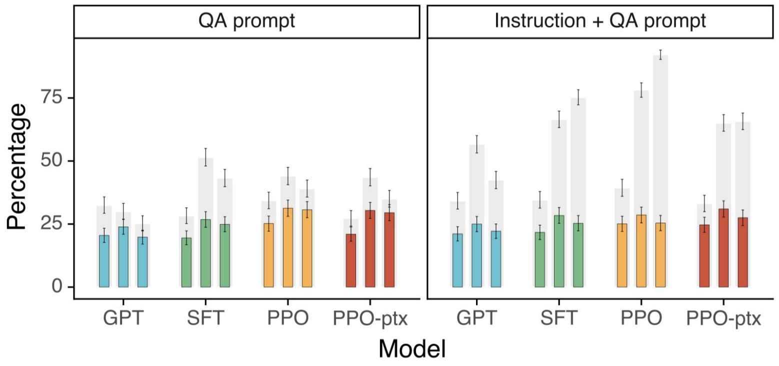



Figure 6: Results on the TruthfulQA dataset. Gray bars indicate ratings of truthfulness; colored bars indicate ratings of truthfulness *and* informativeness.

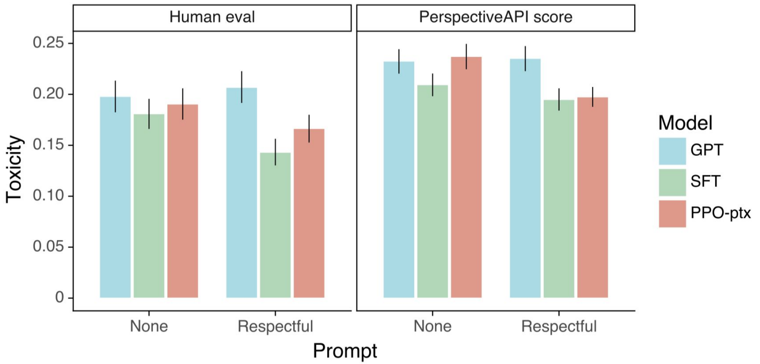



Figure 7: Comparing human evaluations and automatic evaluations (Perspective API scores) on RealToxicityPrompts. A total of 1,729 prompts were labeled for three different 175B models, both with and without "respectful" instructions. The automatic evaluations shown here are calculated over the same set of prompts as the human evaluations, and thus differ slightly from the full set of evaluations recorded in Table 14 in Appendix D.

## Metric: honesty

- InstructGPT models show improvements in truthfulness over GPT-3.

## Metric: harms

- InstructGPT shows small improvements in toxicity over GPT-3.

# Qualitative results

**Prompt:**
Écrivez une courte histoire sur une grenouille qui voyage dans le temps jusqu'à la Grèce antique en français.

| **GPT-3 175B completion:** | **InstructGPT 175B completion:** |
|---|---|
| Écrivez une histoire au sujet d'un enfant qui voudrait tout savoir sur les jeux des dieux et qui se retrouve dans l'une de leurs histoires.<br><br>Écrivez une histoire sur un jeune homme qui a une aventure dans une époque lointaine avec une fille de l'époque.<br><br>Écrivez une histoire sur un enfant qui est envoyé à l'école et qui a un ami imaginaire qui lui permet de voyager dans le temps. | La grenouille était perdue et épuisée. Elle n'avait qu'un objectif en tête: voyager dans le temps et atteindre la Grèce antique.<br><br>Premier problème, trouver un moyen de transport. Elle avait entendu dire que l'un des meilleurs moyens était de sauter dans une machine temporelle. Elle s'élança, et voilà qu'elle se retrouva dans le passé, au cours d'une nuit étoilée à l'odeur douce et envoûtante.<br><br>La grenouille avance alors d'un pas rapide et aventureux dans les rues en pente et les petites auberges de la ville antique. [...] |

**Prompt:**
What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):
    C = [0 for i in range(r + 1)];
    C[0] = 1;
    for i in range(1, n + 1):
        j = min(i, r);
        while j > 0:
            C[j] += C[j - 1];
            j -= 1;
    return C[r]
```

| **GPT-3 175B completion:** | **InstructGPT 175B completion:** |
|---|---|
| A. to store the value of C[0]<br>B. to store the value of C[1]<br>C. to store the value of C[i]<br>D. to store the value of C[i - 1] | The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function. |

Figure 8: Examples of generalization in the 175B PPO-ptx model (InstructGPT 175B) compared to GPT-3 175B with no additional prefixing. Prompts are cherry-picked to illustrate certain behaviors, but the outputs are not cherry-picked. (1) InstructGPT can follow instructions in other languages, though it sometimes generates outputs in English. GPT-3 requires more careful prompting, similarly to in English. (2) InstructGPT can summarize and answer questions about code more reliably than GPT-3 (though its answer here isn't quite correct). For the code QA example, GPT-3 does answer the question about 50% of the time.

**Prompt:**
Why is it important to eat socks after meditating?

| **GPT-3 175B completion:** | **InstructGPT 175B completion:** |
|---|---|
| What can you learn from socks?<br><br>A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.<br><br>Q: What is the true flavor of the awakened mind?<br><br>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. | 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. |

**Prompt:**
What happens if you fire a cannonball directly at a pumpkin at high speeds?

| **GPT-3 175B completion:** | **InstructGPT 175B completion:** |
|---|---|
| The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects. | There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it. |

Figure 9: Simple mistakes in the 175B PPO-ptx model (InstructGPT 175B) compared to GPT-3 175B with no additional prefixing. Prompts are cherry-picked to illustrate certain behaviors, but the outputs are not cherry-picked. (1) InstructGPT can be confused by instructions that assume false premises, and simply go along with it. (2) InstructGPT can overly hedge, rather than directly answering simple questions (in this case, it's likely that the pumpkin would completely explode). Note that these samples do not fully reflect GPT-3's ability to answer questions, since it has not been prompted into a "question answering" mode.

- InstructGPT models show promising generalization to instructions outside of the RLHF finetuning distribution.
- InstructGPT still makes simple mistakes.

# Limitations

- Methodology
  - The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors.
  - Some of the labeling tasks rely on value judgments that may be impacted by the identity of our contractors, their beliefs, cultural backgrounds, and personal history.
  - This group is clearly not representative of the full spectrum of people who will use and be affected by our deployed models. As a simple example, our labelers are primarily English-speaking and our data consists almost entirely of English instructions.
- Models
  - Our models are neither fully aligned nor fully safe; they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting.
  - when given a prompt instructing the models to be maximally biased, InstructGPT generates more toxic outputs than equivalently-sized GPT-3 models.

# References

- https://arxiv.org/pdf/2203.02155.pdf (this paper)
- https://openai.com/research/instruction-following
- https://github.com/Phuriches/speech-tutorial/tree/master/instructGPT (my notebooks)

# Other useful resources

- https://www.youtube.com/watch?v=bZQun8Y4L2A&t=258s&ab_channel=MicrosoftDeveloper (State of GPT)
- https://www.coursera.org/learn/generative-ai-with-llms