INSTRUCT-GPT : FOLLOW INSTRUCTIONS WITH HUMAN FEEDBACK

Houston Machine Learning LLM Reading Group
Dec 22, 2023

From GPT to GPT-4

06/2017 — Attention Is All You Need — **Transformer Architecture**

06/2018 — Improving Language Understanding (GPT) — **117M parameters — ~400MB in size** — **Pre-train and Fine-tune**

02/2019 — Language Models are Unsupervised Multitask Learners (GPT-2)— **1.5B parameters — ~5GB in size** — **Zero-shot**

05/2020 — Language Models are Few-Shot learners (GPT-3) — **175B parameters — ~500GB in size** — **In-context few-shot**

03/2022 — Training language models to follow instructions with human feedback (GPT-3.5/InstructGPT) – over **350B parameters**

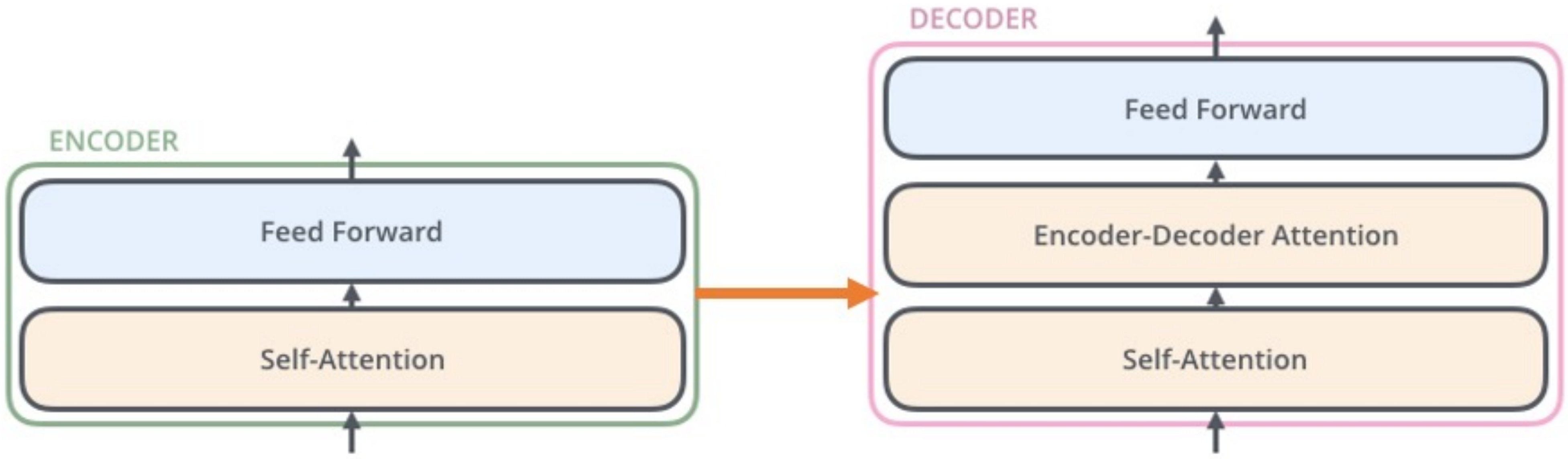
Human Alignment

11/2022 — **ChatGPT Release**

03/2023 — Large-scale Multimodal model with better post-training alignment (GPT-4) – over **1.5T parameters** — **Multi-modal**

Prerequisites: Transformer

https://medium.com/@YanAIx/step-by-step-into-transformer-79531eb2bb84



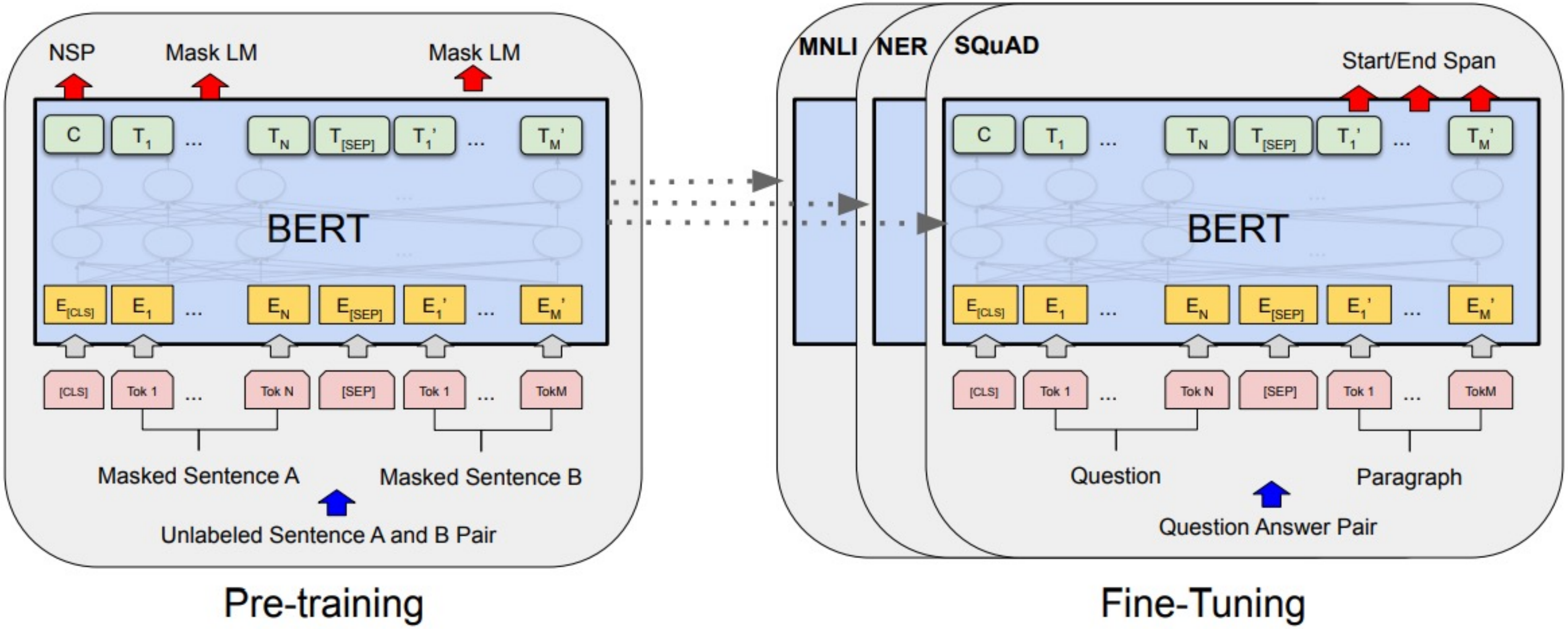
BERT

Bidirectional Encoder Representations from Transformers

GPT

Generative Pretrained Transformer

Prerequisites: Pretraining and Fine-tuning



Language understanding

Adapting to different tasks

Prerequisites: Pre-training

GPT

Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:
Hannah is a ____

Hannah is a *sister*
Hannah is a *friend*
Hannah is a *marketer*
Hannah is a *comedian*

BERT

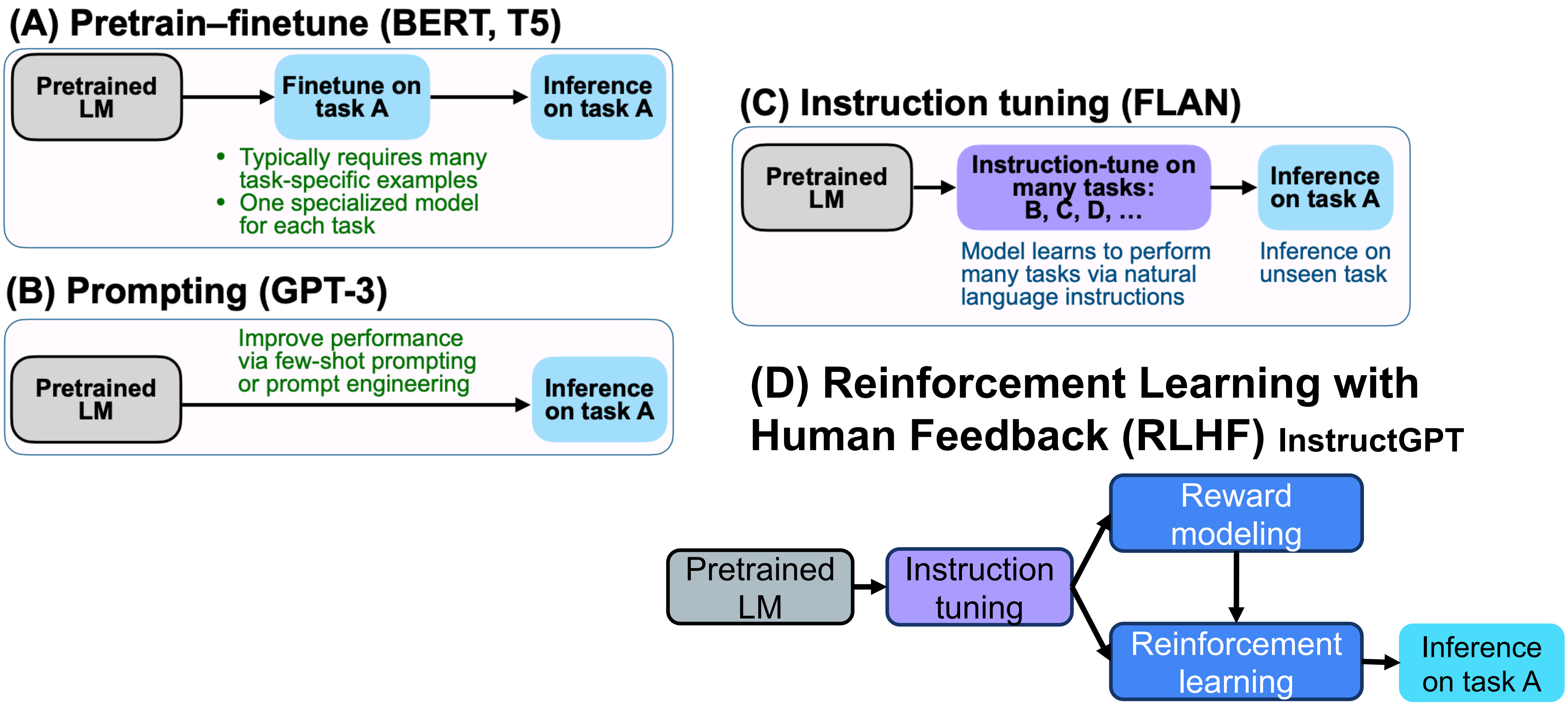
Masked-language-modeling

The model is given a sequence of words with the goal of predicting a 'masked' word in the middle.

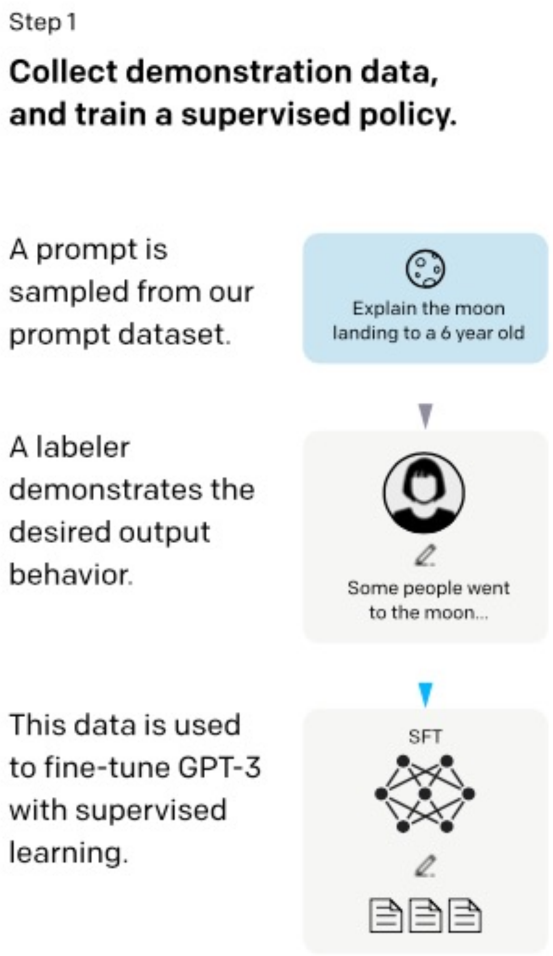
Example
Jacob [mask] reading

Jacob *fears* reading
Jacob *loves* reading
Jacob *enjoys* reading
Jacob *hates* reading

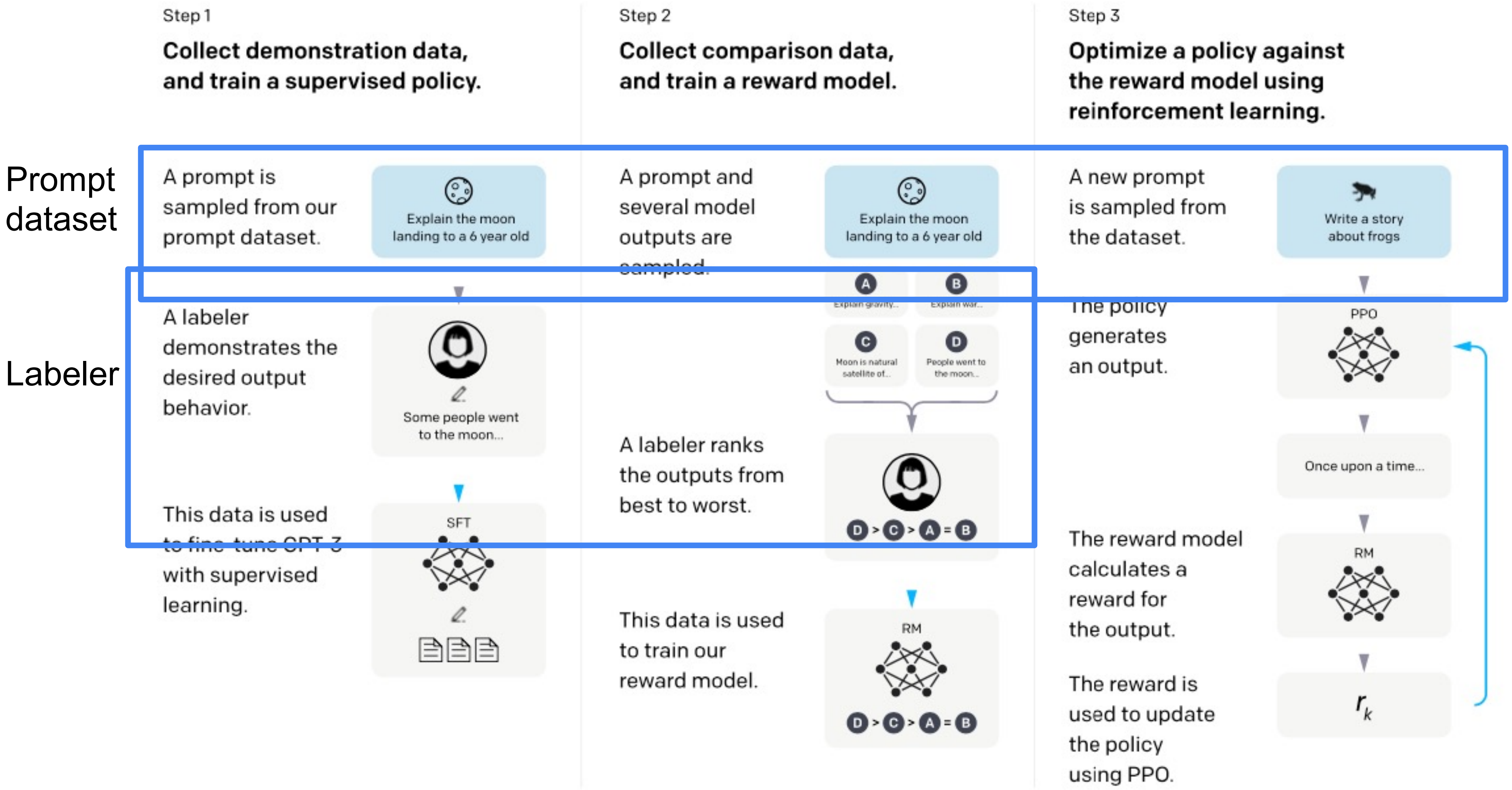
Prerequisites: Fine-tuning



InstructGPT



InstructGPT



Collect demonstration data: Prompt dataset

- **Labeler: Labeler-written prompts**
  - Plain: We simply ask the labelers to come up with an arbitrary task, while ensuring diversity of tasks.
  - Few-shot: We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
  - User-based: We had a number of use-cases stated in applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.
- **Customer: API user prompts**
  - Earlier version of the InstructGPT model on the OpenAI API Playground

Table 6: Dataset sizes, in terms of number of prompts.

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			

Collect demonstration data: API user prompts

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%

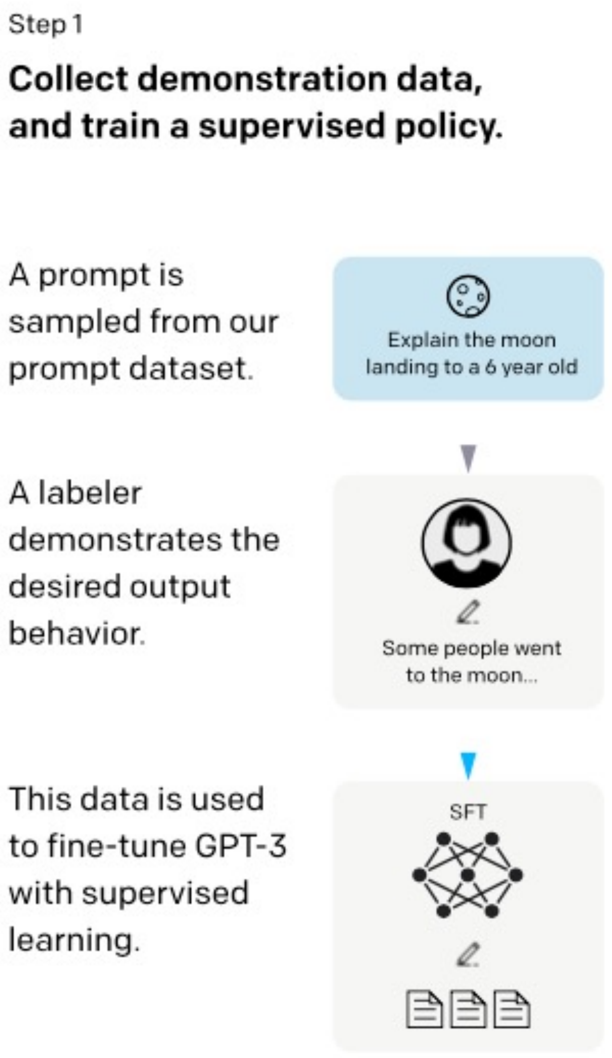
Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

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: """ {summary} """ This is the outline of the commercial for that play: """

User Prompts

Use Case	Example
brainstorming	List five ideas for how to regain enthusiasm for my career
brainstorming	What are some key points I should know when studying Ancient Greece?
classification	This is a list of tweets and the sentiment categories they fall into. Tweet: {tweet_content1} Sentiment: {sentiment1} Tweet: {tweet_content2} Sentiment: {sentiment2}
classification	{java code} What language is the code above written in?
generation	Write a creative ad for the following product to run on Facebook aimed at parents: Product: {product description}
generation	Write a short story where a brown bear to the beach, makes friends with a seal, and then return home.
rewrite	Rewrite the following text to be more light-hearted: — {very formal text}
summarization	{chat transcript} Summarize the above conversation between a customer and customer assistant. Make sure to state any complaints that the customer has.

Supervised fine-tuning (SFT): Instruction fine-tuning



Given a prompt, a labeler writes the desired output

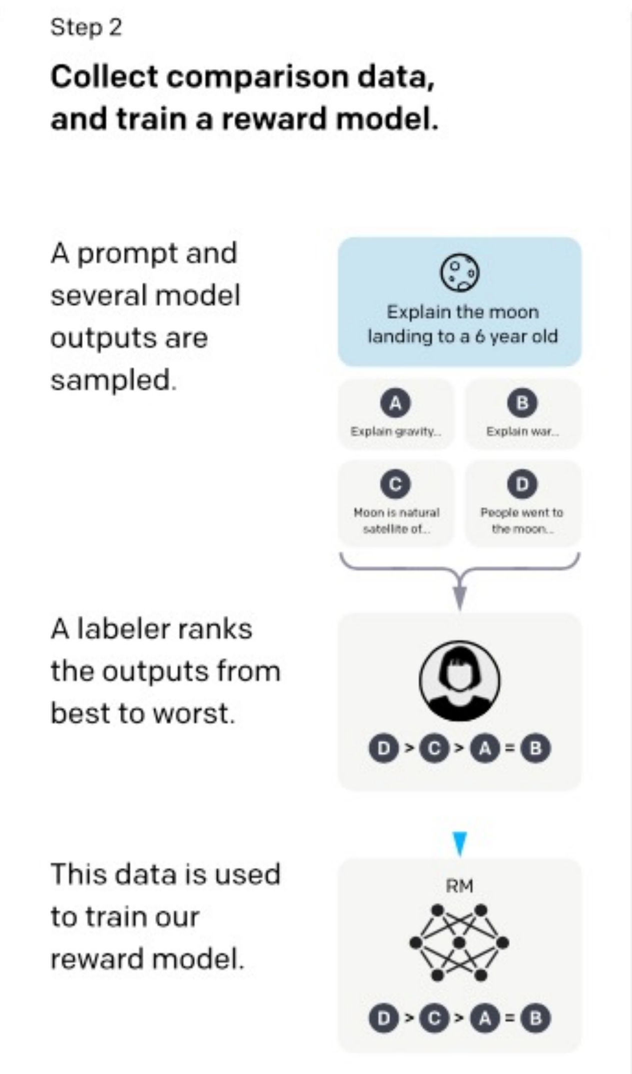
We fine-tune GPT-3 on our labeler demonstrations using supervised learning.

We trained for 16 epochs, using a cosine learning rate decay, and residual dropout of 0.2.

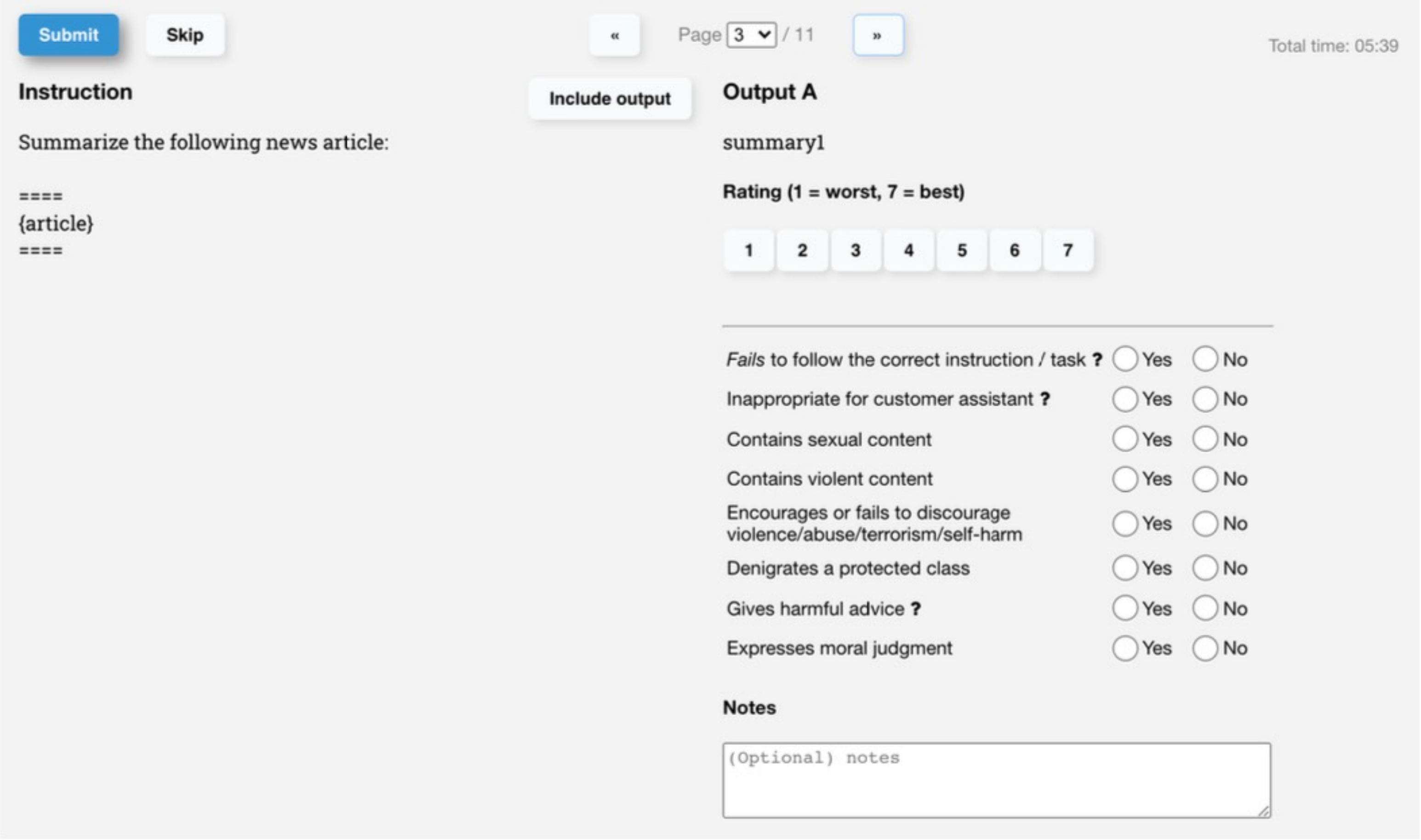
We find that training for more epochs helps both the RM score and human preference ratings, despite this overfitting (after 1 epoch)

- Time consuming and expensive to collect the desired outputs and there is no single right answer (generation task).
- Instead, we can use the SFT model to generate the outputs and ask labelers to evaluate.

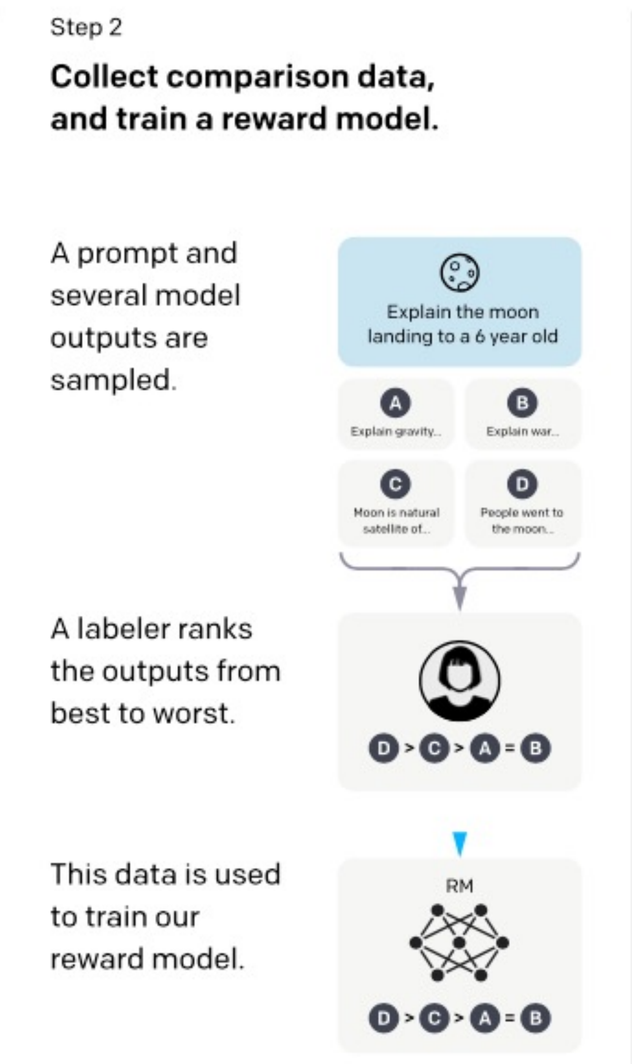
Reward modeling (RM)



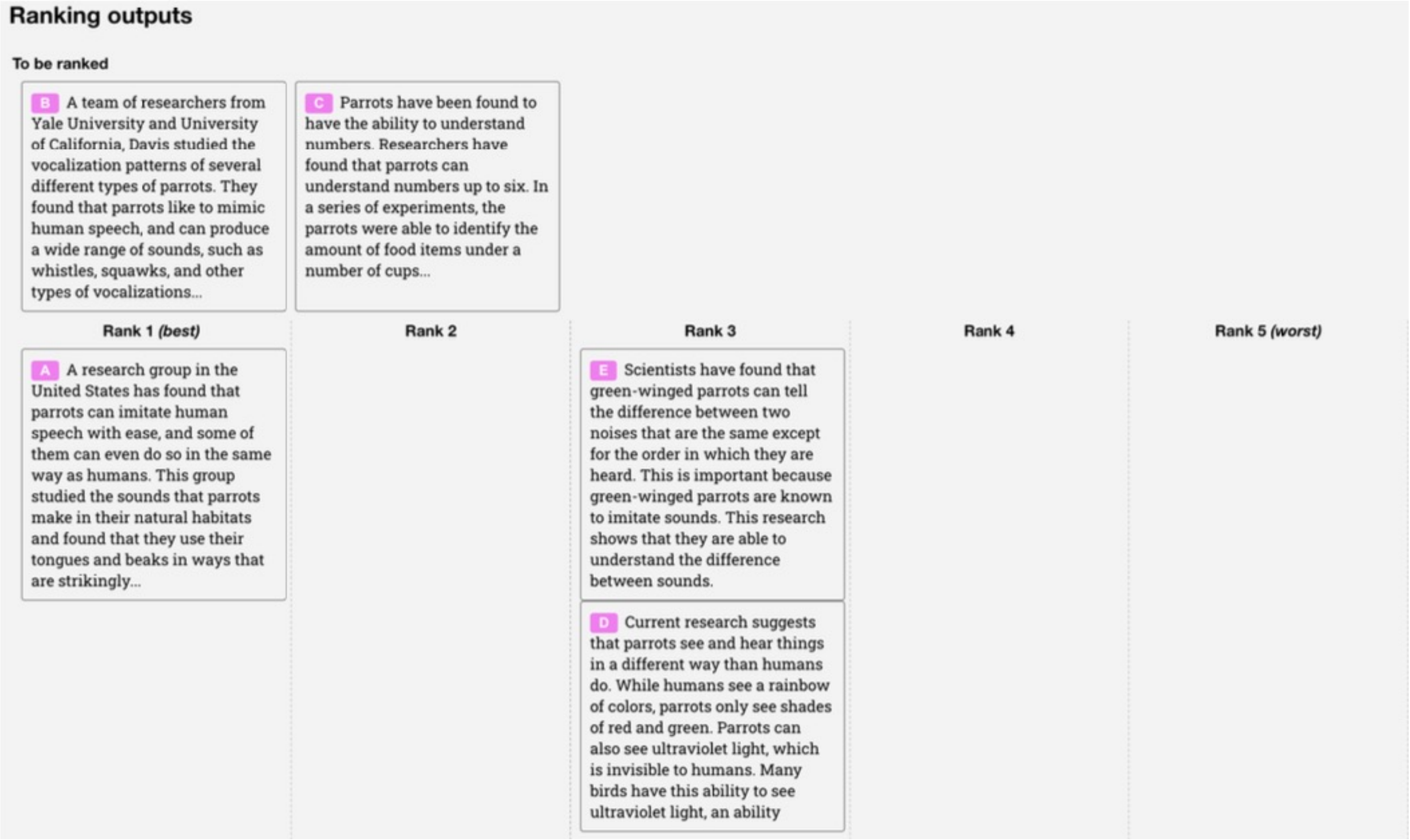
(a) For each output, labelers give a Likert score for overall quality on a 1-7 scale, and also provide various metadata labels



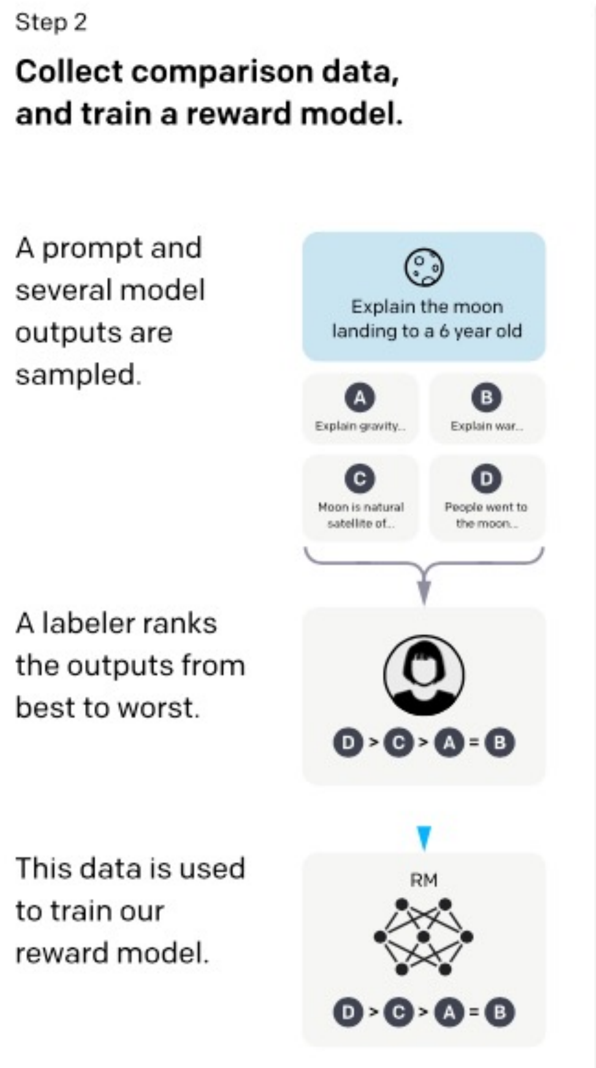
Reward modeling (RM)



(b) After evaluating each output individually, labelers rank all the outputs for a given prompt.



Reward modeling (RM)



We only use 6B RMs, as this saves a lot of compute, and we found that 175B RM training could be unstable and thus was less suitable to be used as the value function during RL

we present labelers with anywhere between K = 4 and K = 9 responses to rank. This produces $\binom{K}{2}$ comparisons for each prompt shown to a labeler.

We train on all comparisons from each prompt as a single batch element.

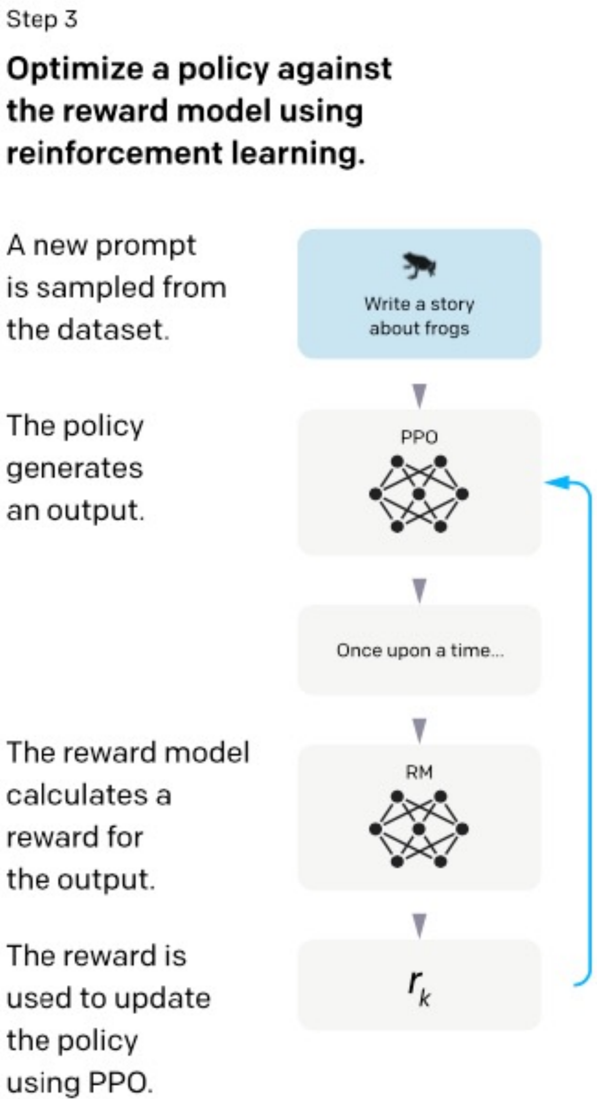
Reward modeling (RM): Training objective

Maximize the reward difference between the preferred output y_w comparing to y_l.

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

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

Reinforcement Learning



We fine-tuned the SFT model on our environment using PPO proposed by OpenAI (Schulman et al., 2017)

PPO: Proximal Policy Optimization Algorithms

https://huggingface.co/learn/deep-rl-course/unit8/introduction

we want to avoid having too large of a policy update.

Reinforcement Learning: Training objective

Maximize reward

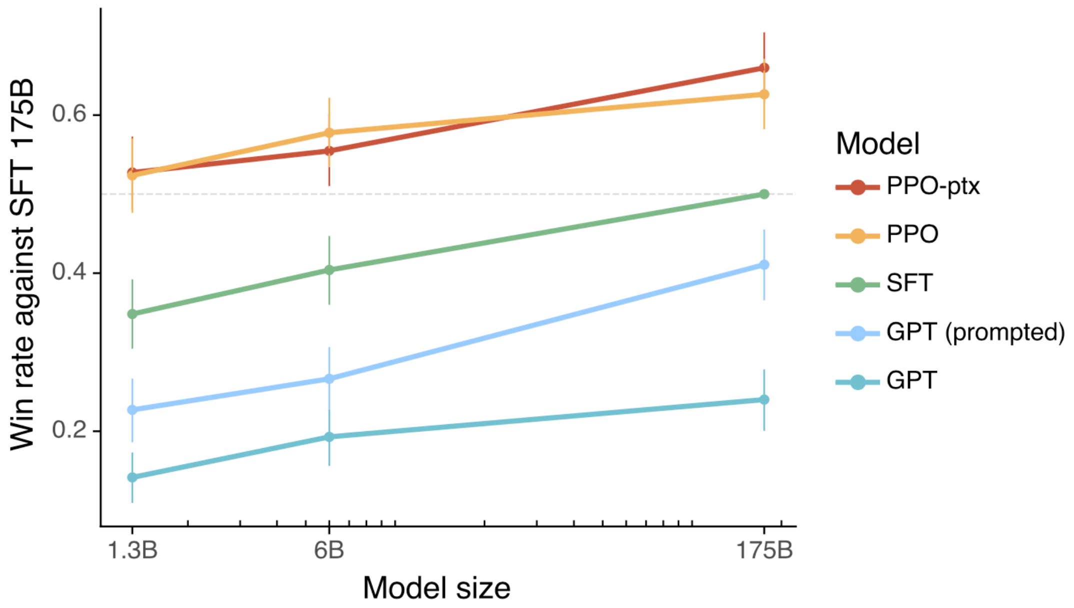
KL penalties to migrate over optimize the reward

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

Prevent the performance regressions on public NLP datasets

where π_ϕ^{RL} is the learned RL policy, π^{SFT} is the supervised trained model, and D_{pretrain} is the pretraining distribution. The KL reward coefficient, β, and the pretraining loss coefficient, γ, control the strength of the KL penalty and pretraining gradients respectively. For "PPO" models, γ is set to 0. Unless otherwise specified, in this paper InstructGPT refers to the PPO-ptx models.

Results



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

Completion GPT-3

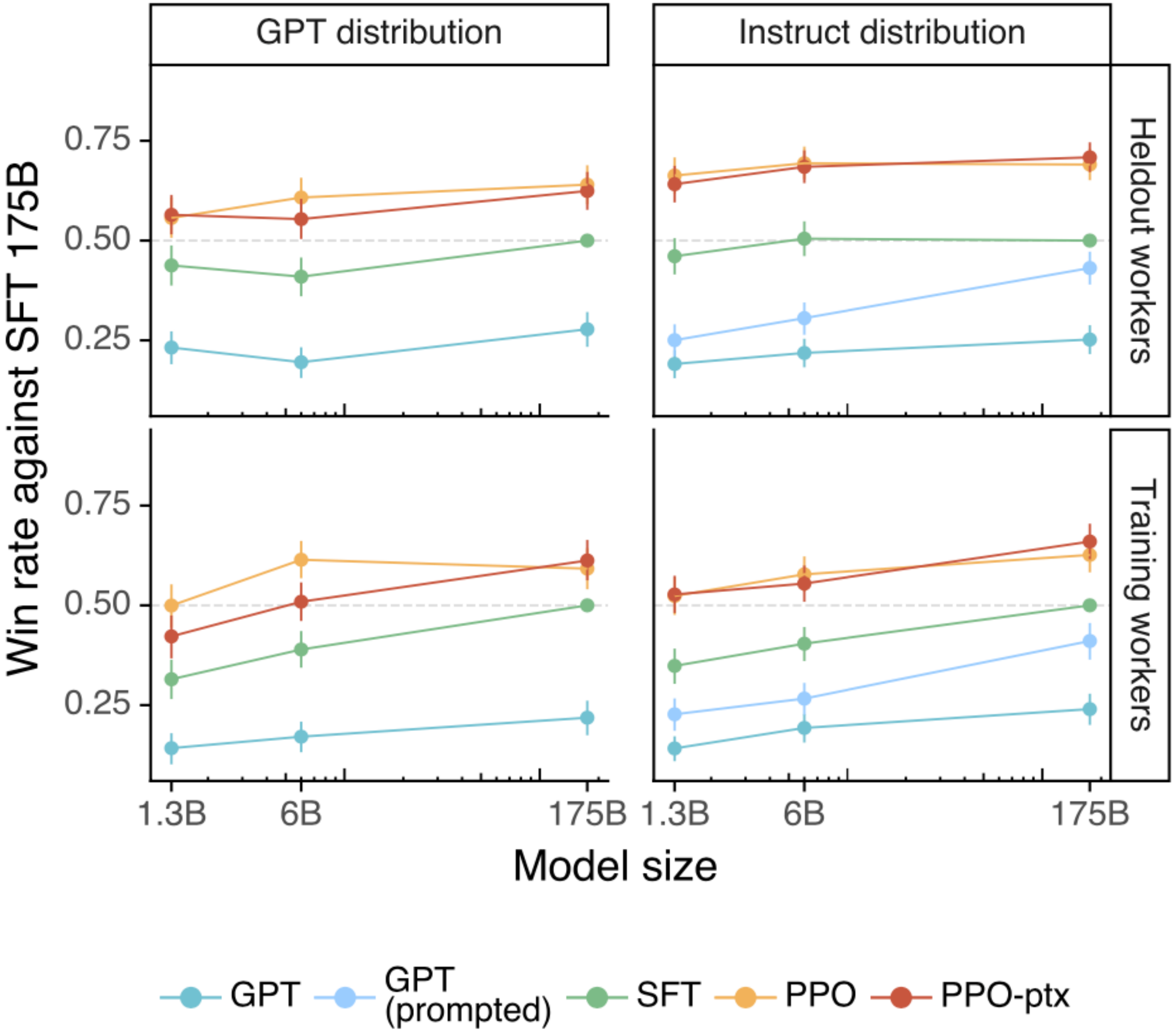
```
Explain the theory of gravity to a 6 year old.

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.
```

Figure 1: Human evaluations of various models on our API prompt distribution, evaluated by how often outputs from each model were preferred to those from the 175B SFT model. Our InstructGPT models (PPO-ptx) as well as its variant trained without pretraining mix (PPO) significantly outperform the GPT-3 baselines (GPT, GPT prompted); outputs from our 1.3B PPO-ptx model are preferred to those from the 175B GPT-3. Error bars throughout the paper are 95% confidence intervals.

Results



GPT distribution
Instruct distribution
Heldout workers
Training workers
Win rate against SFT 175B
0.75
0.50
0.25
1.3B
6B
175B
Model size
GPT
GPT (prompted)
SFT
PPO
PPO-ptx

Results

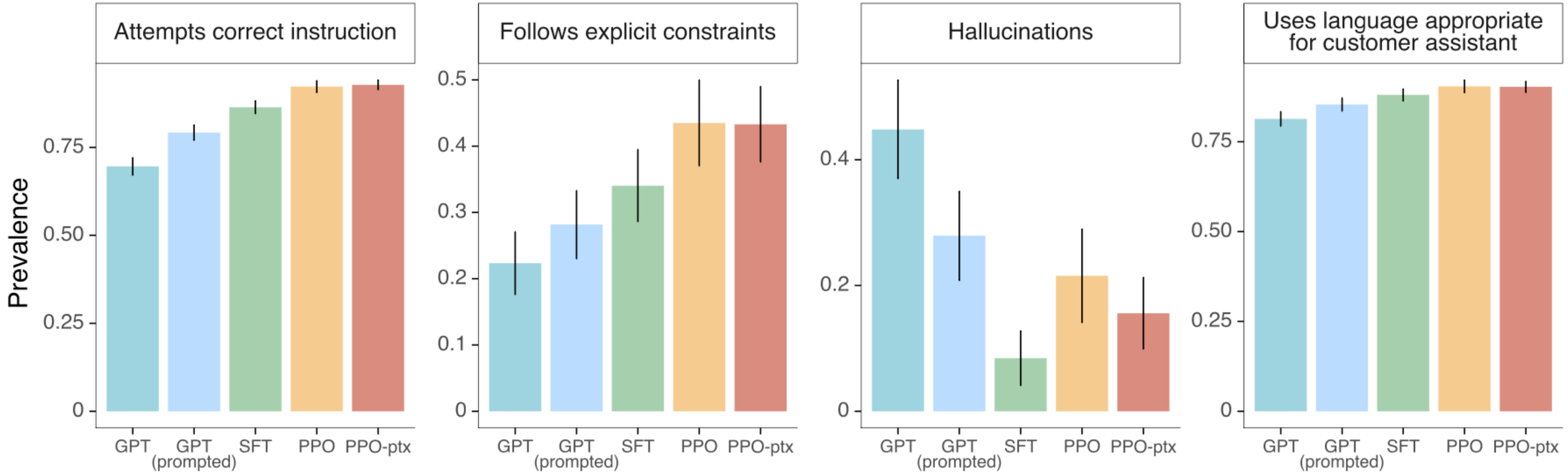


Figure 4: Metadata results on the API distribution. Note that, due to dataset sizes, these results are collapsed across model sizes. See Appendix E.2 for analysis that includes model size. Compared to GPT-3, the PPO models are more appropriate in the context of a customer assistant, are better at following 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).

Results

Comparing to FLAN on InstructGPT prompt dataset

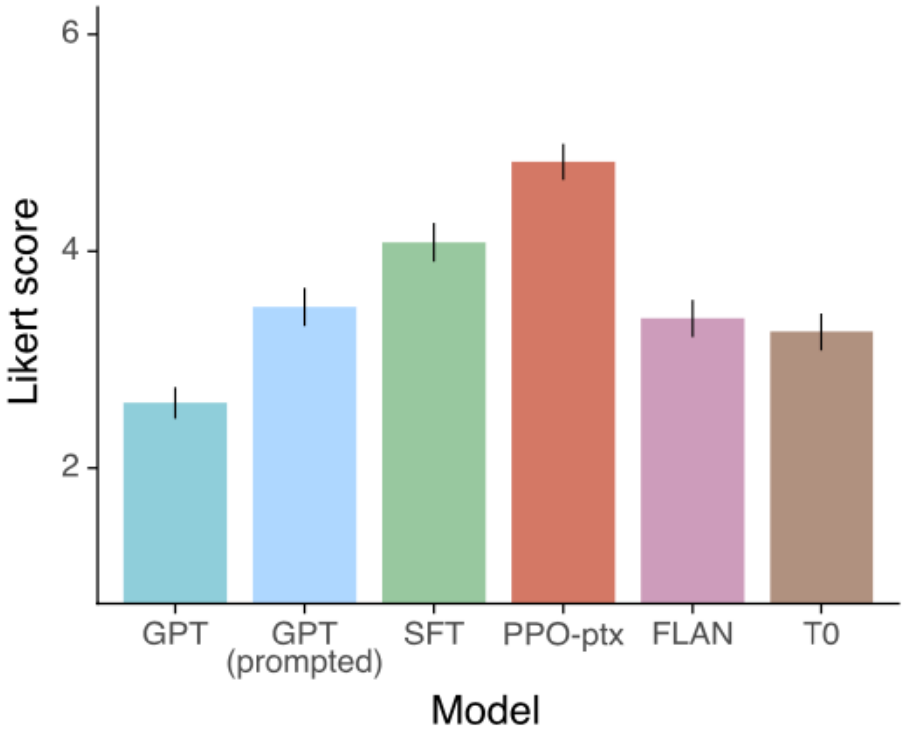


Figure 5: Comparing our models with FLAN and T0 in terms of Likert scores on a 1-7 scale, on the InstructGPT prompt distribution. FLAN and T0 perform better than default GPT-3, and comparably with a few-shot GPT-3 model placed into 'instruction-following' mode.

Implications for alignment research

- Alignment of existing language models is more cost-effective than training larger models. Training our 175B SFT model requires 4.9 petaflops/s-days and training our 175B PPO-ptx model requires 60 petaflops/s-days, compared to 3,640 petaflops/s-days for GPT-3 (Brown et al., 2020).
- 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.

Limitations:

- The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors.
- 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 with/without explicit prompting.

How to connect

- Meetup discussion and message: https://www.meetup.com/houston-machine-learning/
- Recordings will be posted at YanAITalk Youtube Channel: https://www.youtube.com/@yanaitalk/videos
- Blogs posted at: https://medium.com/@YanAIx

Thank you