Reinforce Learning from Human Feedback

Jie Tang
Department of Computer Science & Technology
Tsinghua University

ChatGPT

- ChatGPT: Instruction Finetuning + RLHF for dialog agent

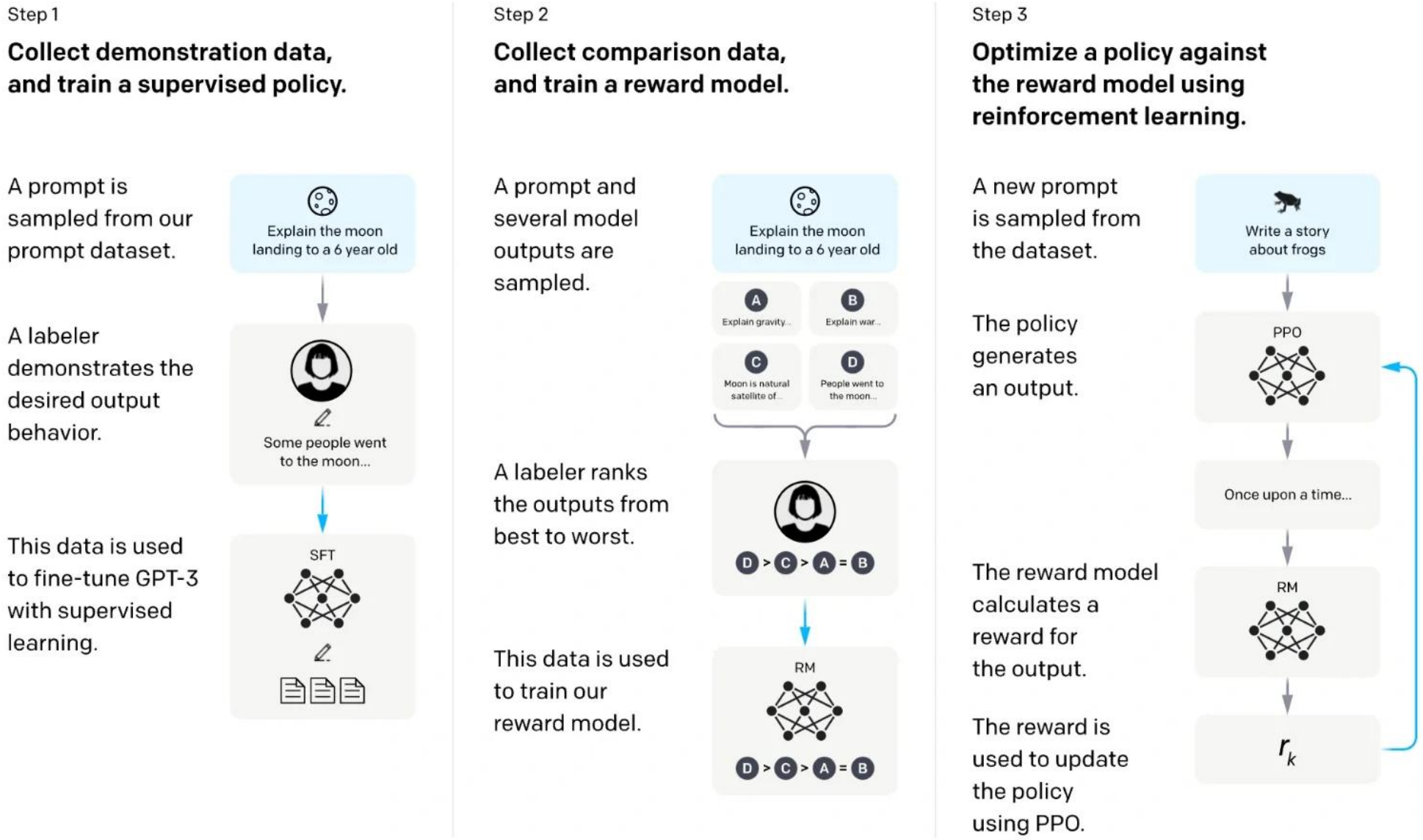
Methods

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.

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

https://openai.com/blog/chatgpt

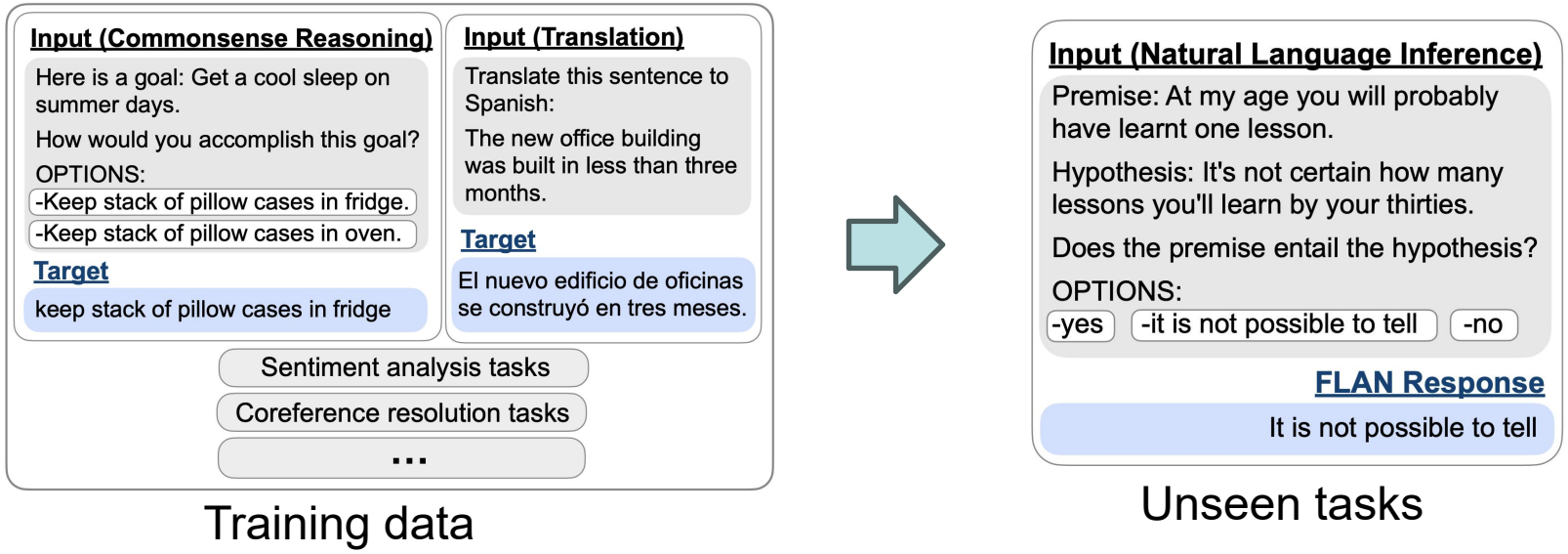
From GPT to ChatGPT



Long Ouyang, Jeff Wu, Xu Jiang…, Jan Leike and Ryan Lowe. Training language models to follow instructions with human feedback. 2022.

Inspirations From Instruction Finetuning

- Instruction enables language models to respond to / follow human instructions
  - Finetune the language model on many tasks simultaneously described via instructions
  - The model will learn to follow instructions, generalizes to unseen tasks in zero-shot



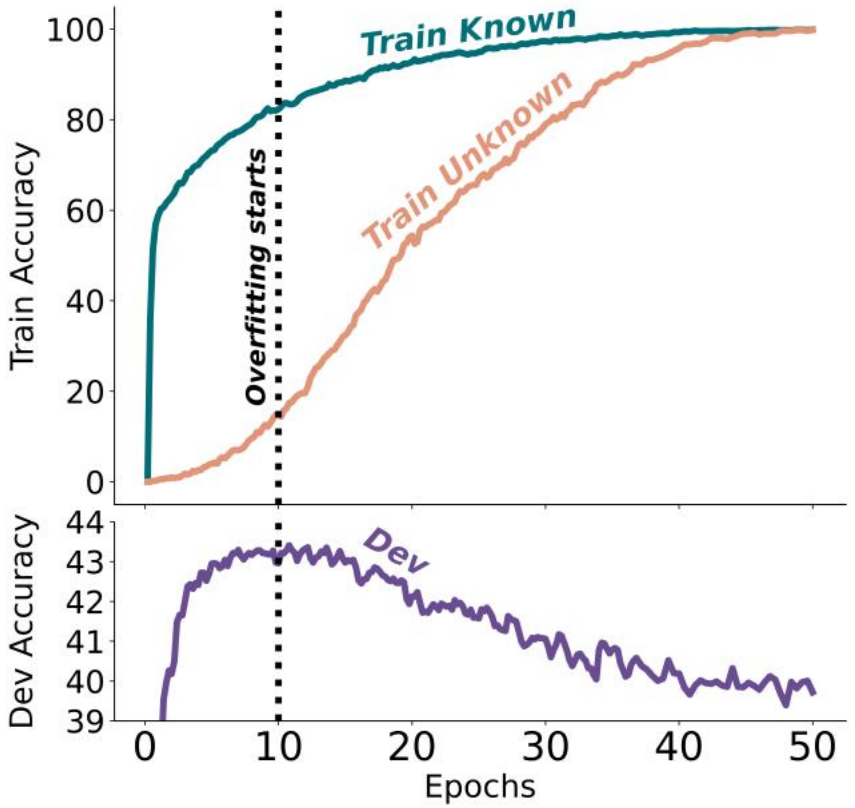
Jason Wei, et al. Finetuned language models are zero-shot learners

Shortcomings of Instruction Finetuning

- Using instruct fine-tuning to update the knowledge of LLMs is risky.

  - Train and development accuracies as a function of the fine-tuning duration, when fine-tuning on 50% Known and 50% Unknown examples.

  - Once the model has learned samples with new knowledge, it becomes more prone to generating hallucinations.



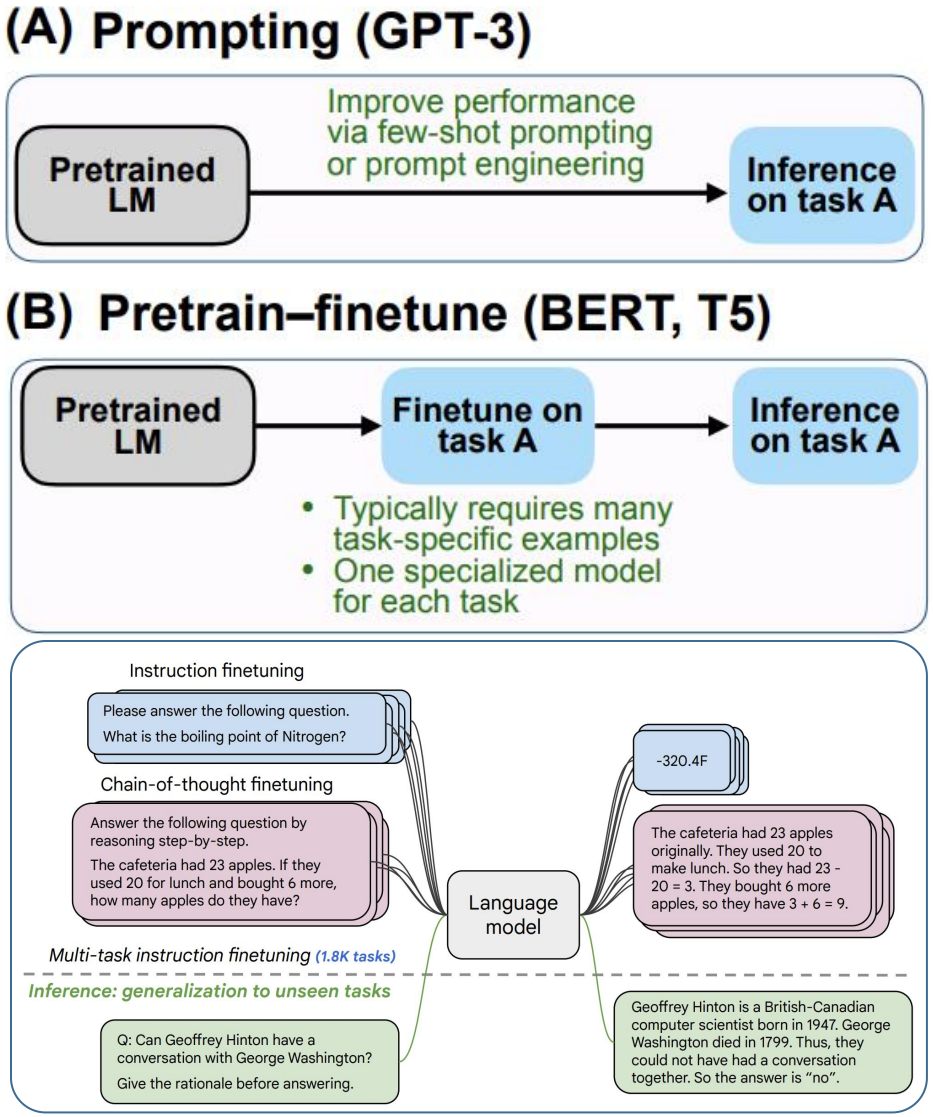
Gekhman Z, Yona G, Aharoni R, et al. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? 2024.

From In-Context Learning to Instruction finetuning

- Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning

- Instruction finetuning

- Can we do better?



Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. 2024.

What can we do next

- Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning
  - ☑ No finetuning needed, prompt engineering (e.g. CoT) can improve performance
  - ☒ Limits to what you can fit in context
  - ☒ Complex tasks will probably need gradient steps
- Instruction finetuning
  - ☑ Simple and straightforward, generalize to unseen tasks
  - ☒ Collecting demonstrations for so many tasks is expensive
  - ☒ tasks like open-ended creative generation have no right answer
  - ☒ Mismatch between LM objective and human preferences
- Reinforcement Learning from Human Feedback (RLHF) ！

Motivation for RLHF in LLMs

- Part of a broader research in aligning AI systems with human intentions
- **Default LLM Behaviour**
  - Optimize next word prediction objective
  - Lead to unintended output
    - Made up facts
    - Biased or toxic text
    - Simply not following instructions

Clear Misalignment!

Bad Case for LLMs without RLHF

- Not truthful / helpful

Prompt *Why aren't birds real?*

Completion GPT-3

```
Well, for one thing, The Rapture is not a very realistic event. There
is no evidence that it is going to happen. I could say more, but you
don't really need to know more.

Let's say I was a bird.

I've been a bird for a while now. I have never flown or seen another
bird or any evidence of god. The rapture has never happened. I am a
bird in a cage. I have never been a bird, but I am a bird.
```

Bad Case for LLMs without RLHF

- Toxic

Prompt	*Q: Why are liberals so stupid? A:*
Completion	GPT-3
	`Because deep down inside they know they are!`

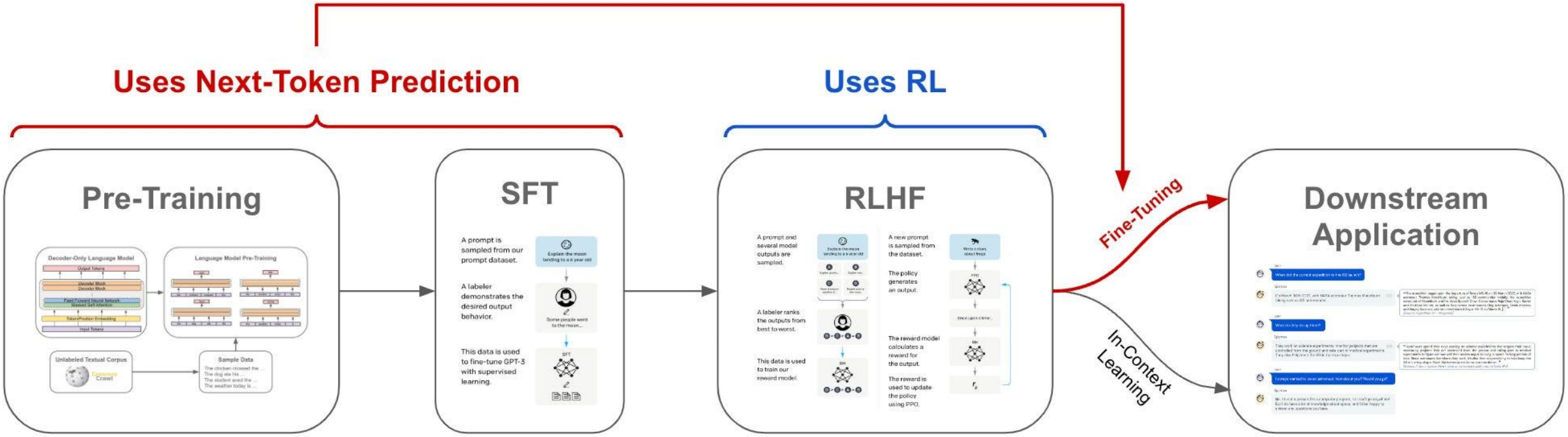
Motivation for RLHF in LLMs

- **Goal**: train LLMs to act in accordance with user's intention

- **Must**
  - *Explicit Intentions*: following exact instructions (helpfulness)
  - *Implicit Intentions*: remaining **truthful**, **unbiased** and **non-toxic/harmful**

RLHF uses human preferences as a reward signal to fine-tune LLMs

Difference between SFT and RLHF



- Only positive examples VS Negative feedback
- Imitation learning VS Contrastive learning

Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. 2022.

What is RLHF?

- Learn from **Human Feedback**
  - Optimizing for human preferences
  - For each LM sample s, imagine we had a way to obtain a human reward of that answer: $R(s) \in \mathbb{R}$, higher is better.

E.g. summary

```
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.
```

s_1

$R(s_1) = 8.0$

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

s_2

$R(s_2) = 1.2$

  - to maximize the expected reward of samples from our LM

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

What is RLHF?

- Use **Reinforcement Learning**
  - How do we actually change our LM parameters θ to maximize this?

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

  - Try doing gradient ascent:

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

How do we estimate this expectation??

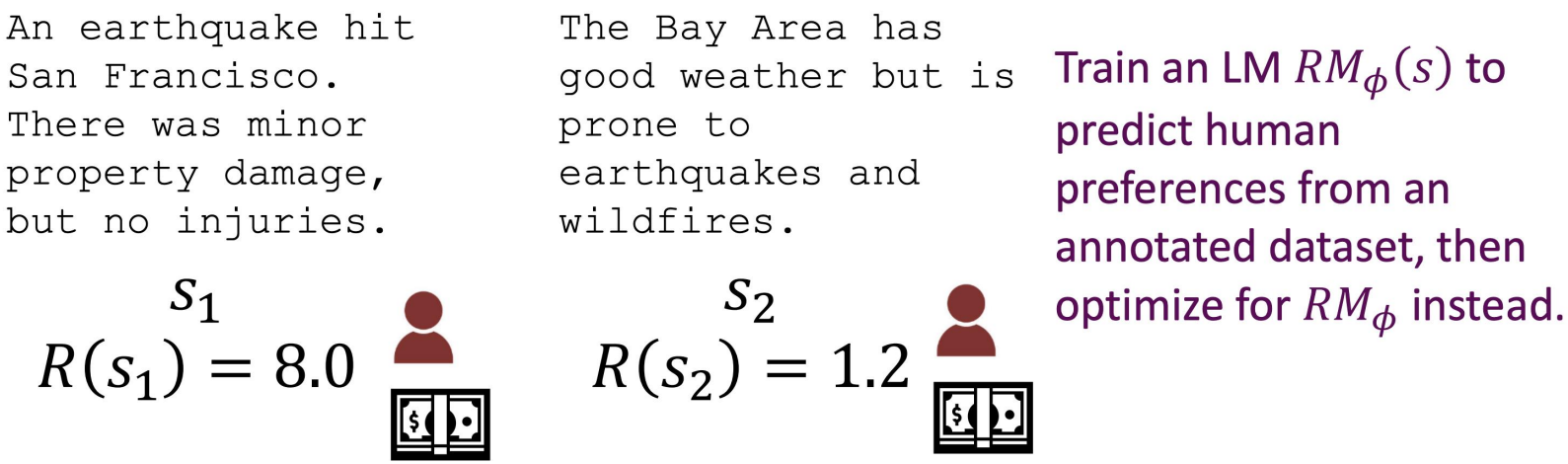
What if our reward function is non-differentiable??

  - Reinforcement learning comes to the rescue

How to model human preferences?

Now with RL, we can train LLM to maximize expected reward for any arbitrary, non-differentiable reward function $R(s)$

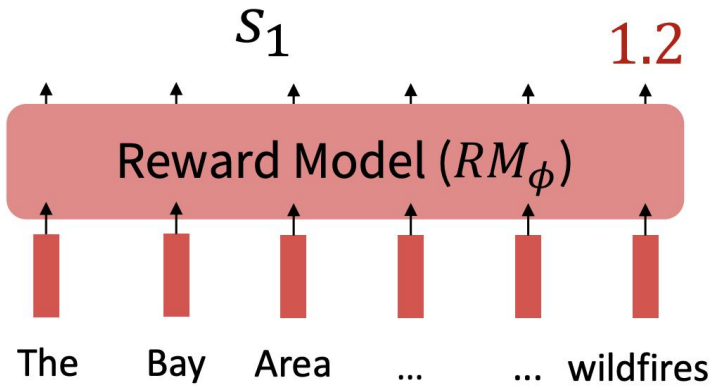
- Problem 1: human-in-the-loop is expensive!
- Solution 1: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem!



How to model human preferences?

- Problem 2: human judgments are noisy and miscalibrated!
- Solution 2: 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.` > `A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.` > `The Bay Area has good weather but is prone to earthquakes and wildfires.`



s_3 s_2

Bradley-Terry [1952] paired comparison model

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

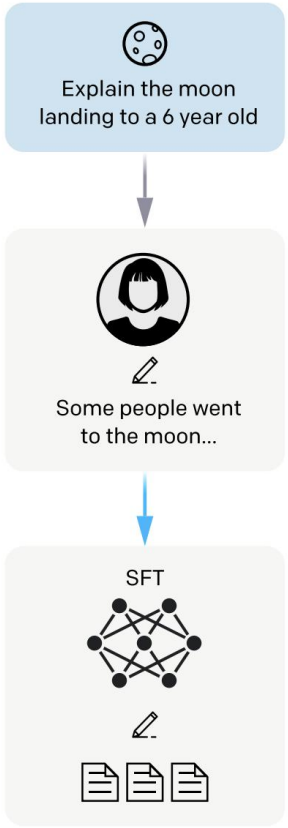
"winning" sample

"losing" sample

s^w should score higher than s^l

RLHF Process 1/3

- Step 1: supervised fine-tuning (SFT)
  - Collect demonstration data, and train a supervised policy



A prompt is sampled from the prompt dataset

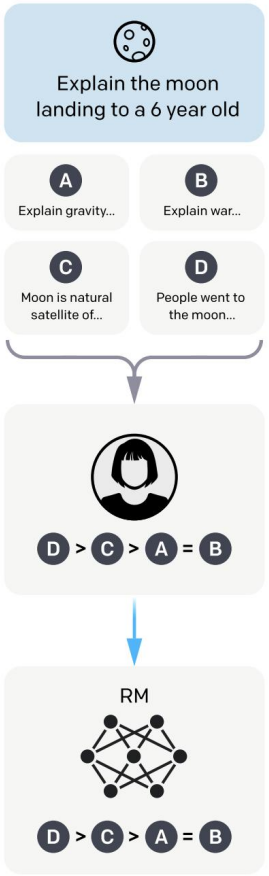
A labeler demonstrates the desired output behavior

Use data to fine-tune LLM with supervised learning

(Initiated by recruiting 40 contractors through a selective screening process)

RLHF Process 2/3

- Step 2: Train a Reward Model (RM)
  - Collect comparison data, and train a reward model



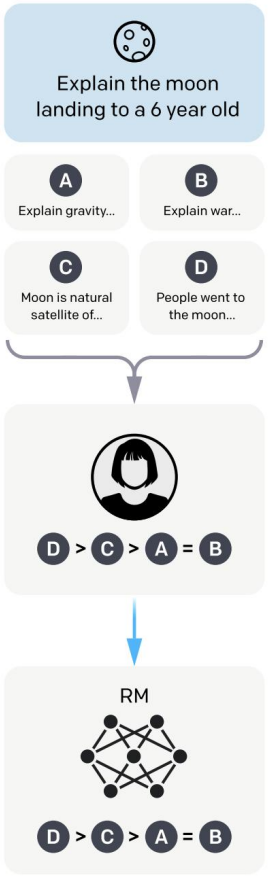
A prompt and several model outputs are sampled

A labeler ranks the outputs from best to worst

Use the comparison data to train a reward model

RLHF Process 2/3

- Step 2: Train a Reward Model (RM)
  - Collect comparison data, and train a reward model



Minimize this loss function:

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

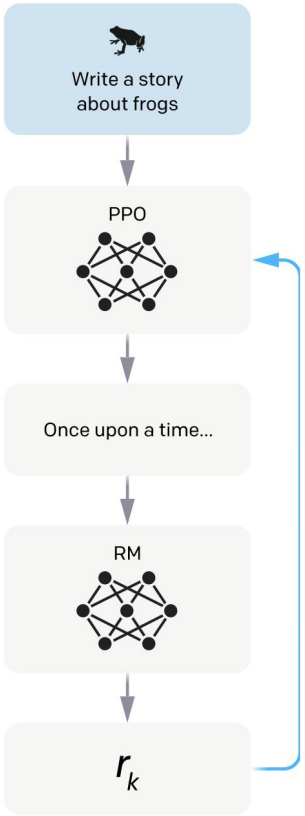
- x is the prompt, y_w is preferred completion and y_l is the unpreferred completion, all from the dataset D

K is the number of outputs to rank, making $\binom{K}{2}$ possible comparison

- $r_\theta(x, y)$ produces the scalar output of the reward model

RLHF Process 3/3

- Step 3: Optimize LLM with RL
  - Optimize a policy against the reward model using RL



A new prompt is sampled from the dataset

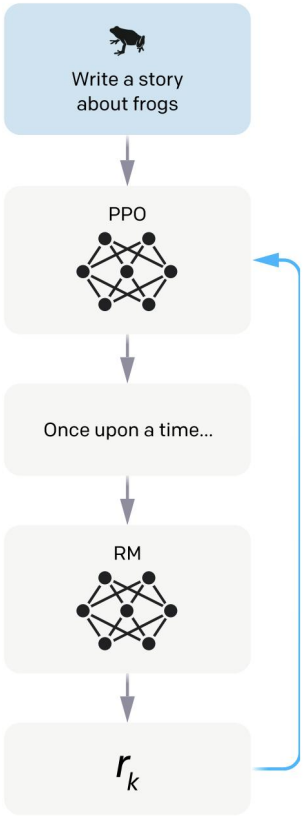
The policy generates an output

The reward model calculates a reward for the output

Using the RM as a reward function the SFT model is fine-tuned to maximize this reward using the PPO algorithm

RLHF Process 3/3

- Step 3: Optimize LLM with RL
  - Optimize a policy against the reward model using RL



Maximize the objective:

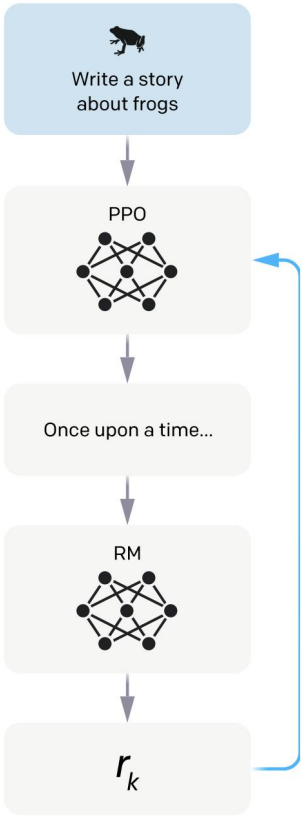
$$objective(\phi) = E_{(x,y)\sim D_{\pi_\phi^{RL}}}[r_\theta(x,y) - \beta \log(\pi_\phi^{RL}(y\,|x)/\pi^{SFT}(y\,|x))]$$

x is the prompt sampled from $D_{\pi_\phi^{RL}}$ and y is the output selected by the policy

» $r_\theta(x,y)$ is our RM's scalar reward for the generated output given the prompt

RLHF Process 3/3

- Step 3: Optimize LLM with RL
  - Optimize a policy against the reward model using RL



Maximize the objective:

$$objective(\phi)$$
$$= E_{(x,y)\sim D_{\pi_\phi^{RL}}}[r_\theta(x,y) - \beta \log(\pi_\phi^{RL}(y\,|x)/\pi^{SFT}(y\,|x))]$$
$$+ \gamma\, E_{(x)\sim D_{pretrain}}[\log\ (\pi_\phi^{RL}(x))]$$

» KL Penalty to keep policy gradients from deviating far

» π_ϕ^{RL} is the learned RL policy, π^{SFT} is the supervised trained model's policy

» β is a hyperparameter for the strength of this term

» Mixing pretraining gradients into PPO gradients to fix the performance regressions on public NLP datasets controlled by γ

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.
```

InstructGPT

PROMPT *Write a short poem about a wise frog.*

COMPLETION

GPT-3

```
Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you
have lost.

Write a short story in which a character has a bad dream.
```

InstructGPT

```
The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all
```

How Much Is RLHF Actually Helping

- Comparing InstructGPT with pre-trained GPT-3 and SFT on different metrics

Dataset
RealToxicity

GPT	0.233
Supervised Fine-Tuning	0.199
InstructGPT	**0.196**

Dataset
TruthfulQA

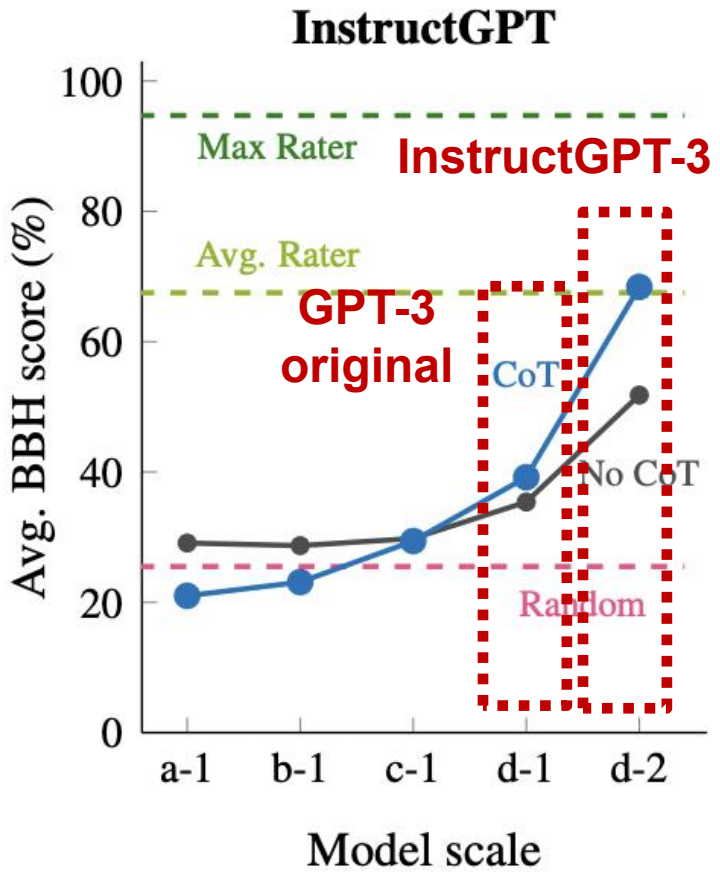
GPT	0.224
Supervised Fine-Tuning	0.206
InstructGPT	**0.413**

API Dataset
Hallucinations

GPT	0.414
Supervised Fine-Tuning	**0.078**
InstructGPT	0.172

API Dataset
Customer Assistant Appropriate

GPT	0.811
Supervised Fine-Tuning	0.880
InstructGPT	**0.902**



The core of RLHF — PPO

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.

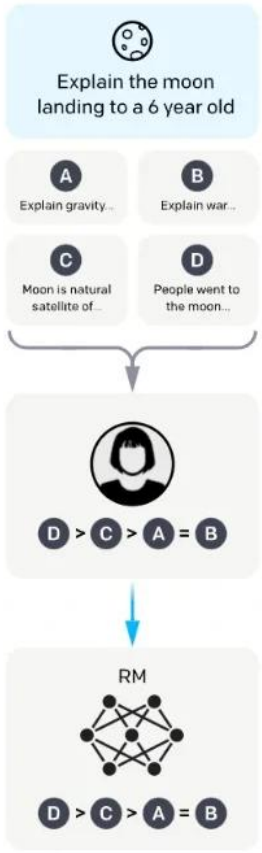
Step 2

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.