education

Лекция 5. InstructGPT

Елисеев Антон
30.10.2024

LLM Recap

Вопрос: Какая архитектура у LLM?

LLM Recap

Вопрос: Как собирать данные?

Instruction following

Paper: https://arxiv.org/abs/2203.02155

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:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
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.

Instruction following

Training objective:

what we **want** vs
what we **told** model to do

What we **told** model to do:

- predict the next token on a webpage from the internet

Alignment $\longrightarrow$

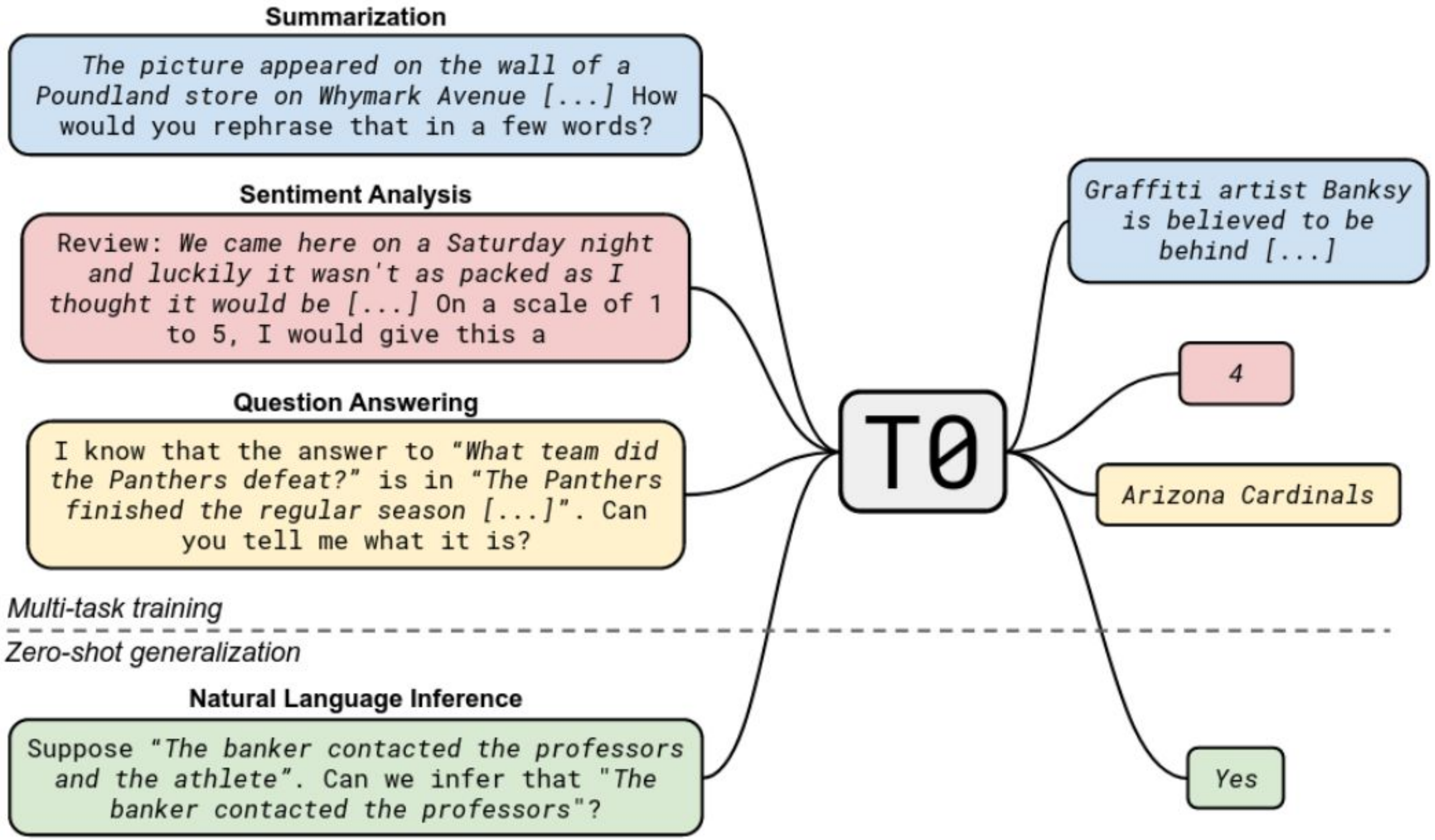
What we want model to do:

- follow the user's instructions helpfully and safely

The language modeling objective is misaligned

Instruction following

Paper: https://arxiv.org/abs/2110.08207



SFT issues

Вопрос: В чем проблемы SFT?

SFT issues

- Supervised seq2seq learning:

$$P\left(y_{t+1} | x, y_{0:t}\right), \qquad y_{0:t} \sim reference$$

- Inference

$$P\left(y_{t+1} | x, \hat{y}_{0:t}\right), \qquad \hat{y}_{0:t} \sim \text{???}$$

SFT issues

- Supervised seq2seq learning:

$$P\left(y_{t+1} \middle| x, y_{0:t}\right), \qquad y_{0:t} \sim reference$$

- Inference

$$P\left(y_{t+1} \middle| x, \hat{y}_{0:t}\right), \qquad \hat{y}_{0:t} \sim model$$

If model ever makes something that isn't in data,
It gets volatile from next time-step!

SFT issues

There's more then one correct translation.

You don't need to learn all of them.

```
Source:  在 找 给 家里 人 的 礼物 .

Versions:
i 'm searching for some gifts for my family.
i want to find something for my family as presents.
i 'm about to buy some presents for my family.
i 'd like to buy my family something as a gift.
i 'm looking for a present for my family.
...
```

SFT issues

There's more then one correct translation.
You don't need to learn all of them.

Source: 在 找 给 家里 人 的 礼物 .

Versions:	Model 1 **p(y|x)**	Model 2 **p(y|x)**
(version 1)	1e-2	0.99
(version 2)	2e-2	1e-100
(version 3)	1e-2	1e-100
(all rubbish)	0.96	0.01

SFT issues

There's more then one correct translation.
You don't need to learn all of them.

Source: 在 找 给 家里 人 的 礼物 .

Versions:	Model 1 **p(y|x)**	Model 2 **p(y|x)**
(version 1)	1e-2	0.99
(version 2)	2e-2	1e-100
(version 3)	1e-2	1e-100
(all rubbish)	0.96	0.01
	better llh 96% rubbish	worse llh 1% rubbish

SFT issues

Вопрос: Почему тогда не использовать только RL?

SFT issues

Вопрос: Почему тогда не использовать только RL?

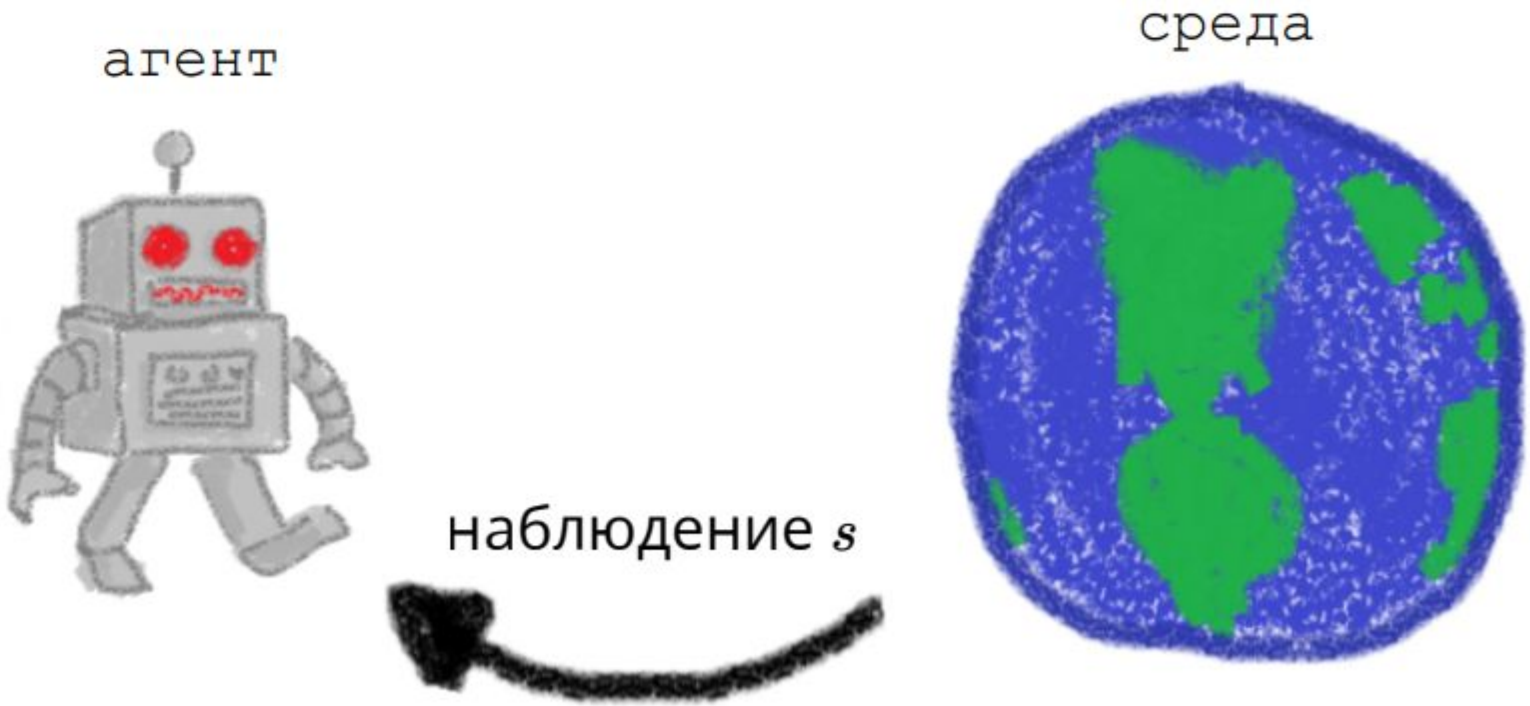
1. Это RL -> плохо сходится, нужно много данных
2. Говорит “плохо”, но не говорит, как правильно

Вывод: делаем SFT, поверх него RL

RL recap

Политика агента:

$$a := \pi(s)$$

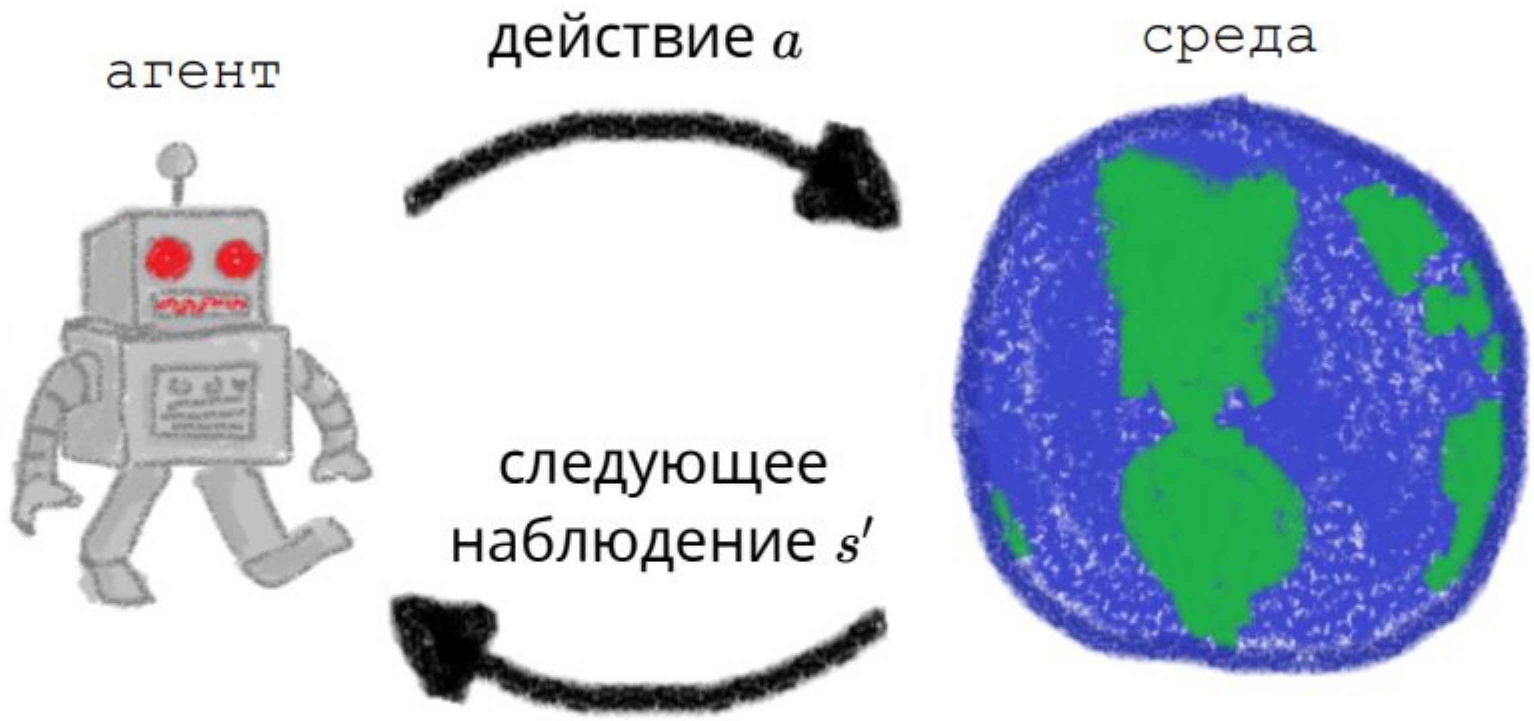


Decision Process - выбор действий по наблюдениям

RL recap

Политика агента:

$$a := \pi(s)$$



Decision Process - выбор действий по наблюдениям

RL recap

Supervised learning:

$$\nabla llh = \underset{x, y_{opt} \sim D}{E} \nabla \log P_\theta(y_{opt}|x)$$

Policy gradient:

$$\nabla J = \underset{\substack{s \sim d(s) \\ a \sim \pi(a|obs(s))}}{E} \nabla \log \pi(a|s) Q(s,a)$$

RL recap

Supervised learning:

$$\nabla llh = \underset{s, a_{opt} \sim D}{E} \nabla \log \pi_\theta(a_{opt} | s)$$

Policy gradient:

$$\nabla J = \underset{\substack{s \sim d(s) \\ a \sim \pi(a|obs(s))}}{E} \nabla \log \pi_\theta(a|s) Q(s,a)$$

RL recap

Supervised learning:

$$\nabla llh = \underset{s, a_{opt} \sim D}{E} \nabla \log \pi_\theta(a_{opt} | s)$$

reference

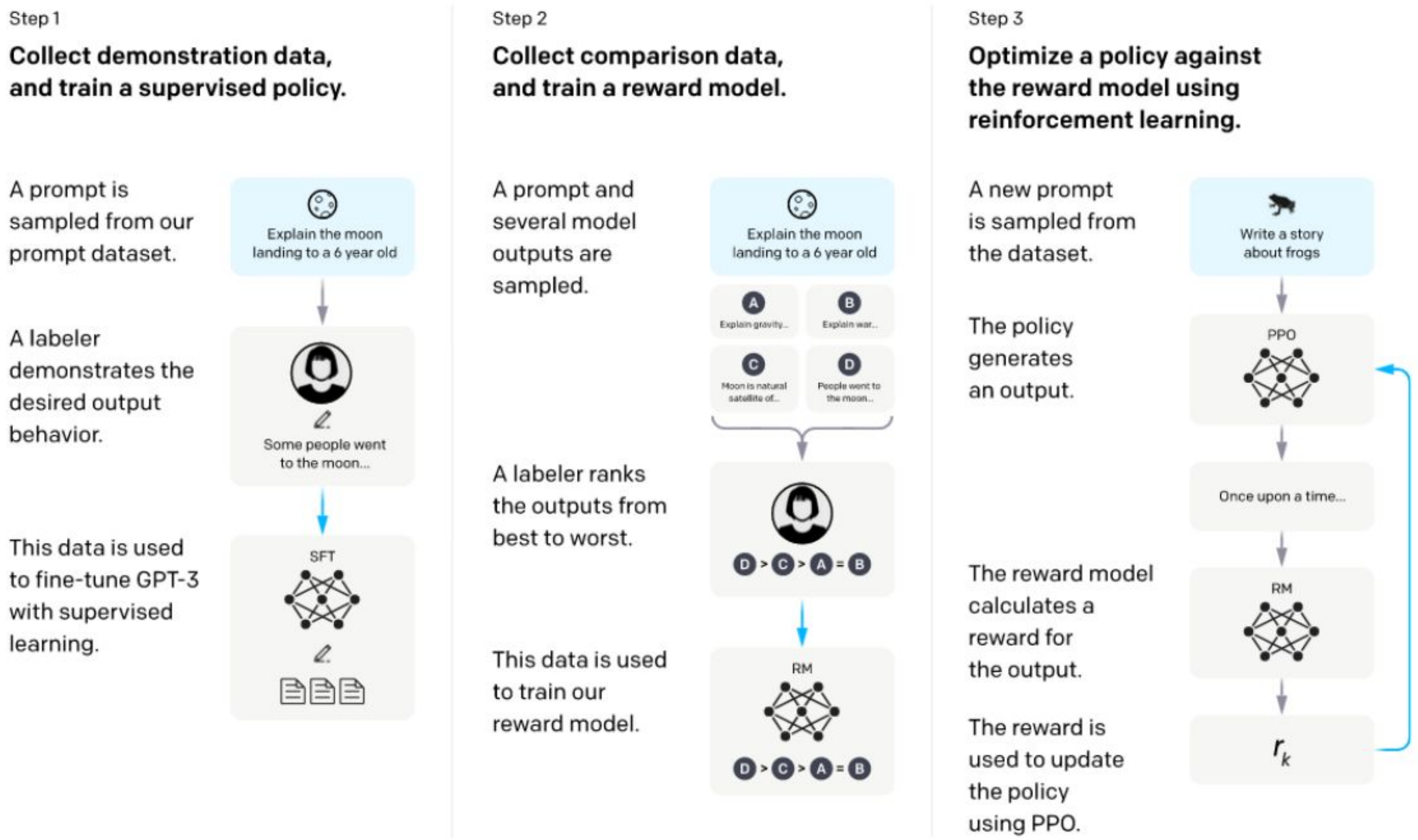
Policy gradient:

$$\nabla J = \underset{\substack{s \sim d(s) \\ a \sim \pi(a | obs(s))}}{E} \nabla \log \pi_\theta(a | s) Q(s, a)$$

generated

RLHF

Paper: https://arxiv.org/abs/2203.02155



RLHF

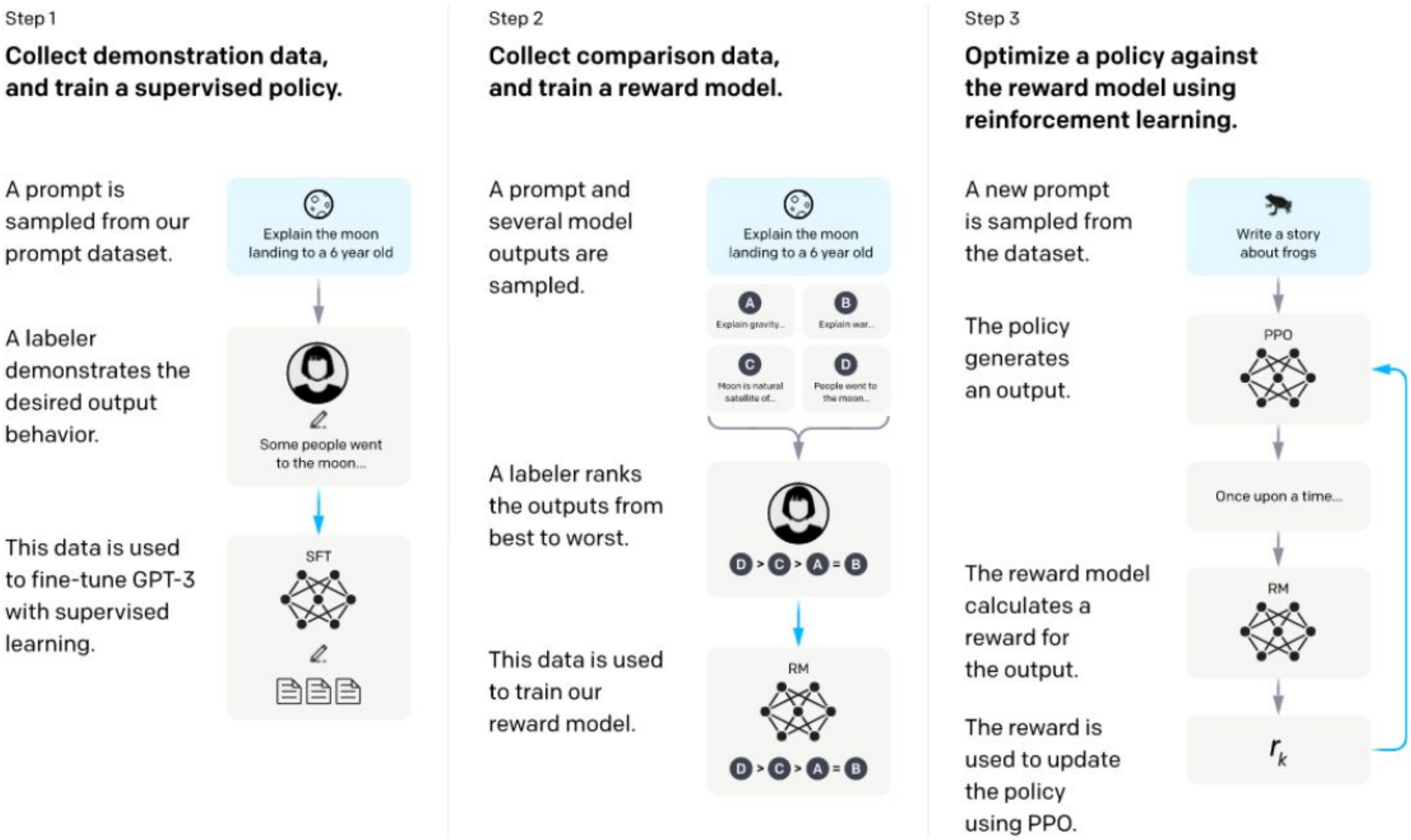
Вопрос: Что плохого в этой схеме?

RLHF

$$\mathbf{E}_{a\sim\pi_\theta(a|s)}\left[r_\psi(s,a) - \beta\mathrm{KL}\big(\pi_\theta(a|s)||\pi_{\mathrm{SFT}}(a|s)\big)\right] \to \max_\theta$$

RLHF

Paper: https://arxiv.org/abs/2203.02155



Reward modeling. Модель Брэдли-Терри

Cons:

1. Не отношения порядка, парадокс Кондорсе

Pros:

1. Делает так же, как и люди

1. Рефлексивность: $a \preccurlyeq a$.
2. Антисимметричность: если $a \preccurlyeq b$ и $b \preccurlyeq a$, то $a = b$.
3. Транзитивность: если $a \preccurlyeq b$ и $b \preccurlyeq c$, то $a \preccurlyeq c$.

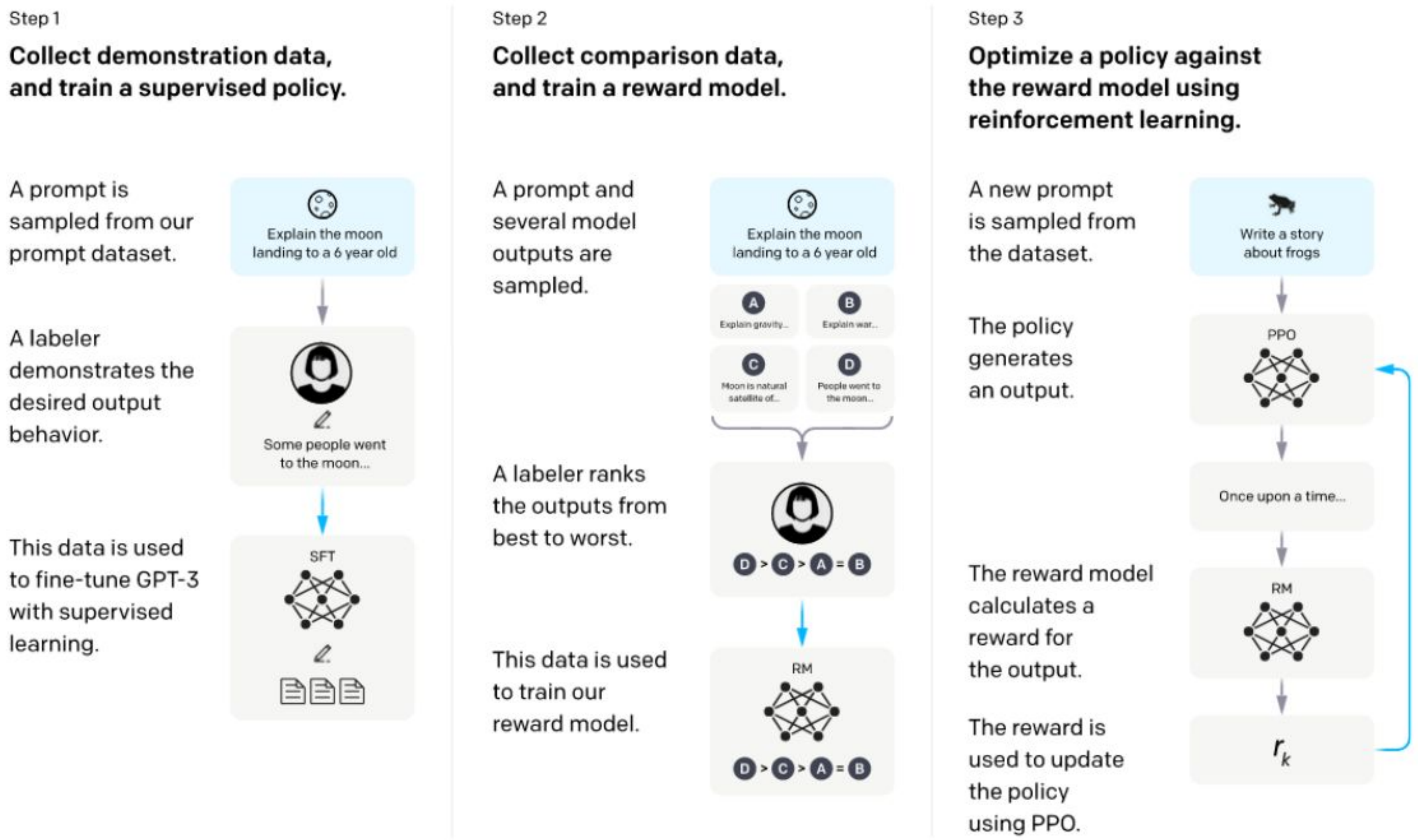
Reward modeling. Модель Брэдли-Терри

$$P(a > b|s) = \sigma(r_\psi(s, a) - r_\psi(s, b)) \qquad \sigma(x) = \frac{1}{1 + \exp(-x)}$$

$$\sum_{(s,winner,loser)\in\mathbf{D}} \log\sigma(r_\psi(s, winner) - r_\psi(s, loser)) \to \max_\psi$$

RLHF

Paper: https://arxiv.org/abs/2203.02155



PPO issues

Вопрос: Какие проблемы у такого подхода?

PPO issues

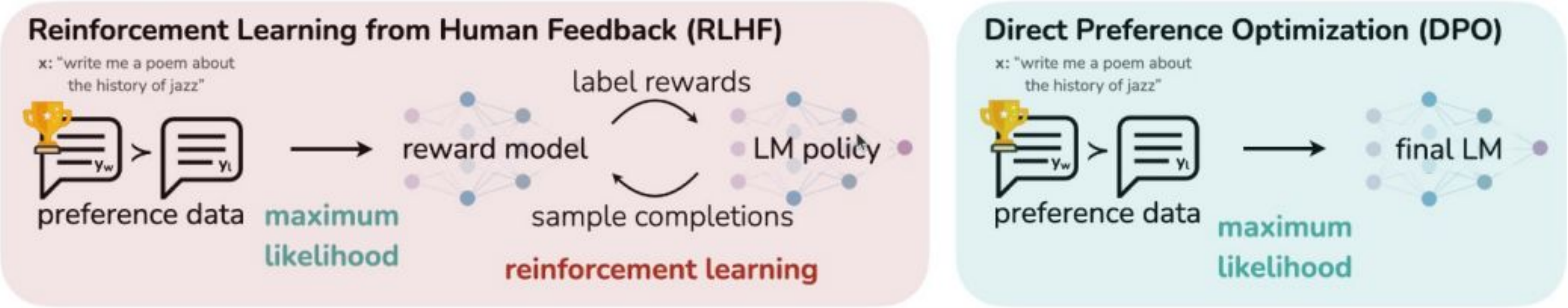
Вопрос: Какие проблемы у такого подхода?

1. Это RL -> плохо сходится, у PPO много гиперпараметров
2. Много GPU памяти
   Пример: gpt3 175B требует 7TB gpu-памяти ~ 88 A100 80G ~ $2M
3. Надо инферить модель на каждом шагу

DPO

Paper: https://arxiv.org/abs/2305.18290

RLHF is complicated, let's make it simpler!



Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their beh
methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align wit
However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the
reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables e
standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call Direct Preference Optimization (DPO), is stable, p
the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human p
exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn

DPO

$$\mathbf{E}_{a\sim\pi_\theta(a|s)}\left[r_\psi(s,a) - \beta\mathrm{KL}\left(\pi_\theta(a|s)||\pi_{\mathrm{SFT}}(a|s)\right)\right] \to \max_\theta$$

DPO

$$\mathbf{E}_{a\sim\pi_\theta(a|s)}\left[r_\psi(s,a) - \beta \mathrm{KL}\left(\pi_\theta(a|s)||\pi_{\mathrm{SFT}}(a|s)\right)\right] \to \max_\theta$$

$$\pi^*(a|s) = \frac{1}{Z(s)}\pi_{\mathrm{SFT}}(a|s)e^{\frac{1}{\beta}r(s,a)}$$

$$Z(s) = \sum_a e^{\frac{1}{\beta}r_{(}s,a)}$$

DPO

$$\pi^*(a|s) = \frac{1}{Z(s)} \pi_{\mathrm{SFT}}(a|s) e^{\frac{1}{\beta} r(s,a)}$$

$$Z(s) = \sum_a e^{\frac{1}{\beta} r_(s,a)}$$

$$r(s,a) = \beta \log \frac{\pi^*(a|s)}{\pi_{\mathrm{SFT}}(a|s)} + \beta \log Z(s)$$

$$r_\theta(s,a) = \beta \log \frac{\pi_\theta(a|s)}{\pi_{\mathrm{SFT}}(a|s)} + \beta \log Z(s)$$

DPO

$$\sum_{(s,winner,loser)\in\mathbf{D}} \log\sigma(r_\theta(s, winner) - r_\theta(s, loser)) \to \max_\theta$$

$$r_\theta(s,a) = \beta\log\frac{\pi_\theta(a|s)}{\pi_{\text{SFT}}(a|s)} + \beta\log Z(s)$$

$$\sum_{(s,winner,loser)\in\mathbf{D}} \log\sigma\Big(\beta\Big[\log\frac{\pi_\theta(winner|s)}{\pi_{\text{SFT}}(winner|s)} - \log\frac{\pi_\theta(loser|s)}{\pi_{\text{SFT}}(loser|s)}\Big]\Big) \to \max_\theta$$

DPO

Плюсы:
1. не RL -> лучше и быстрее сходится
2. Только один гиперпараметр - beta
3. Надо хранить только обучаемую модель
4. Не инферим, только прогоняем существующий ответ

Q&A

Спасибо за внимание!

Елисеев Антон

education