基于人类反馈的强化学习

卢艳峰

August 26, 2023

前言

- 实践基于人类反馈的强化学习（**reinforcement learning from human feed back，RLHF**）

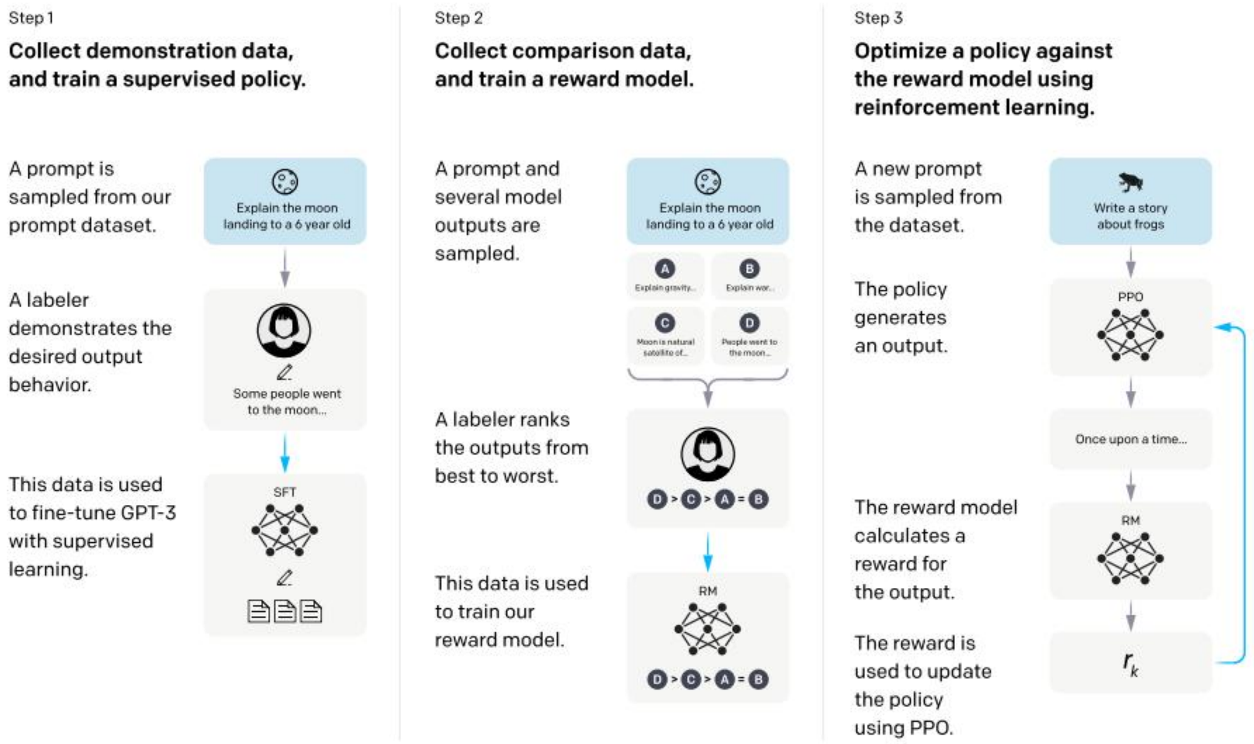
InstructGPT

- 研究目的：将预训练模型（GPT）与用户的需求对齐，使得预训练模型变得有用、诚实和无害。
- 研究前提：一个预训练模型（GPT）。

InstructGPT

- 步骤 1：收集示范数据（如问答数据），训练一个监督策略（神经网络模型）。
- 步骤 2：收集比较数据，训练一个奖励模型。
- 步骤 3：使用近端策略优化算法（Proximal Policy Optimization，PPO）针对奖励模型优化策略。

InstructGPT



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.

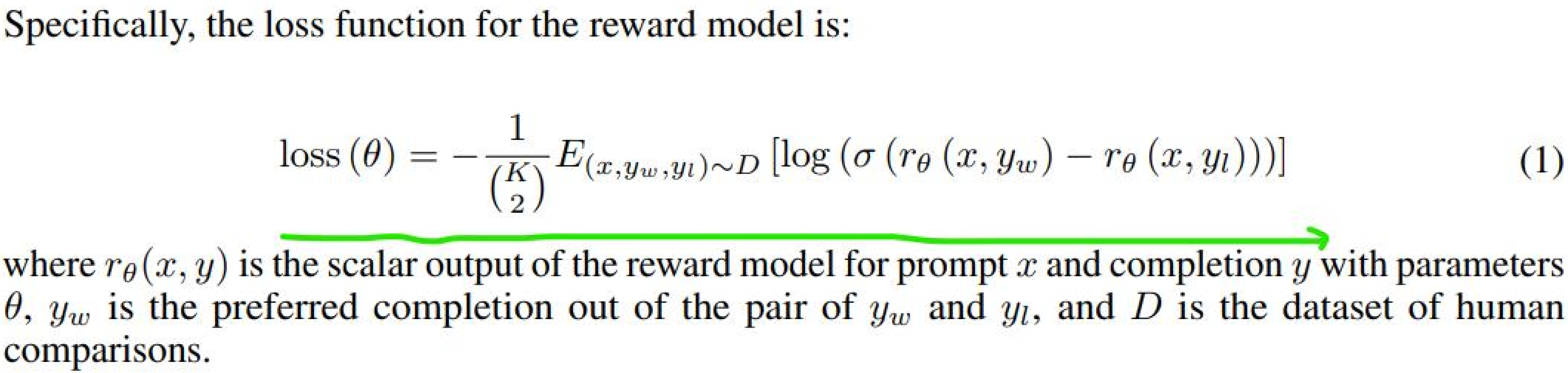
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

InstructGPT

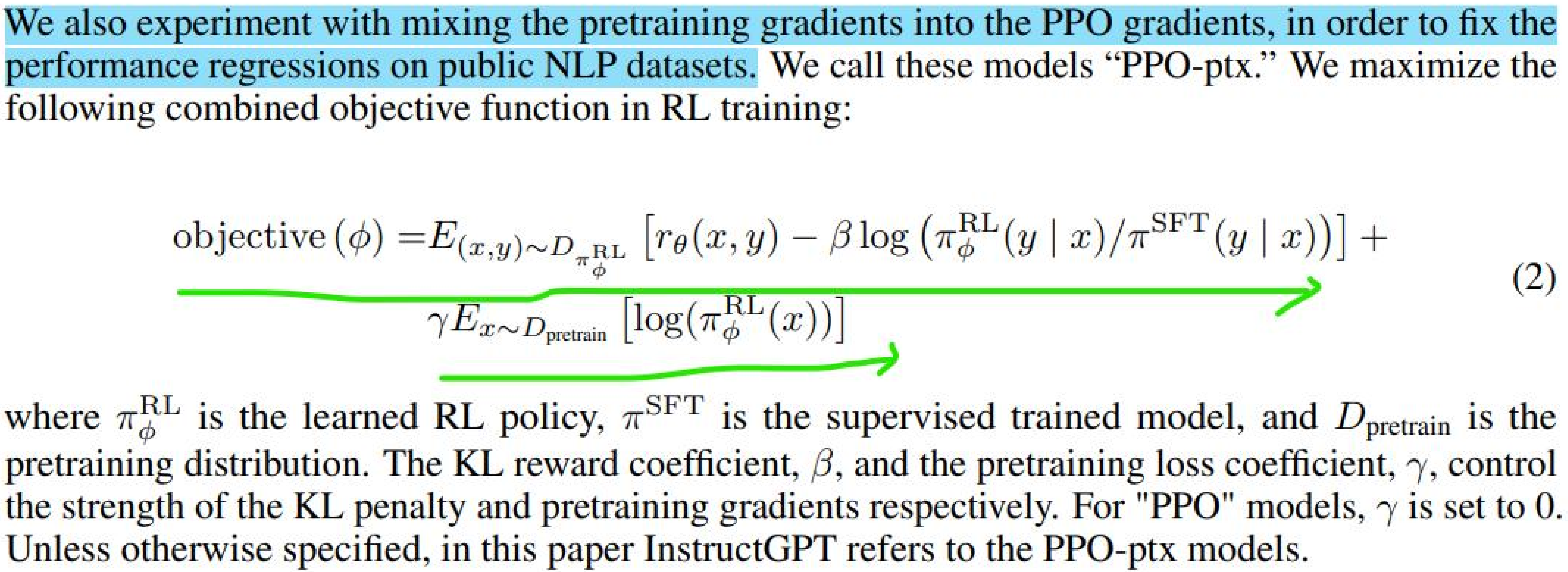


Specifically, the loss function for the reward model is:

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

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.

InstructGPT



We also experiment with mixing the pretraining gradients into the PPO gradients, in order to fix the performance regressions on public NLP datasets. We call these models “PPO-ptx.” We maximize the following combined objective function in RL training:

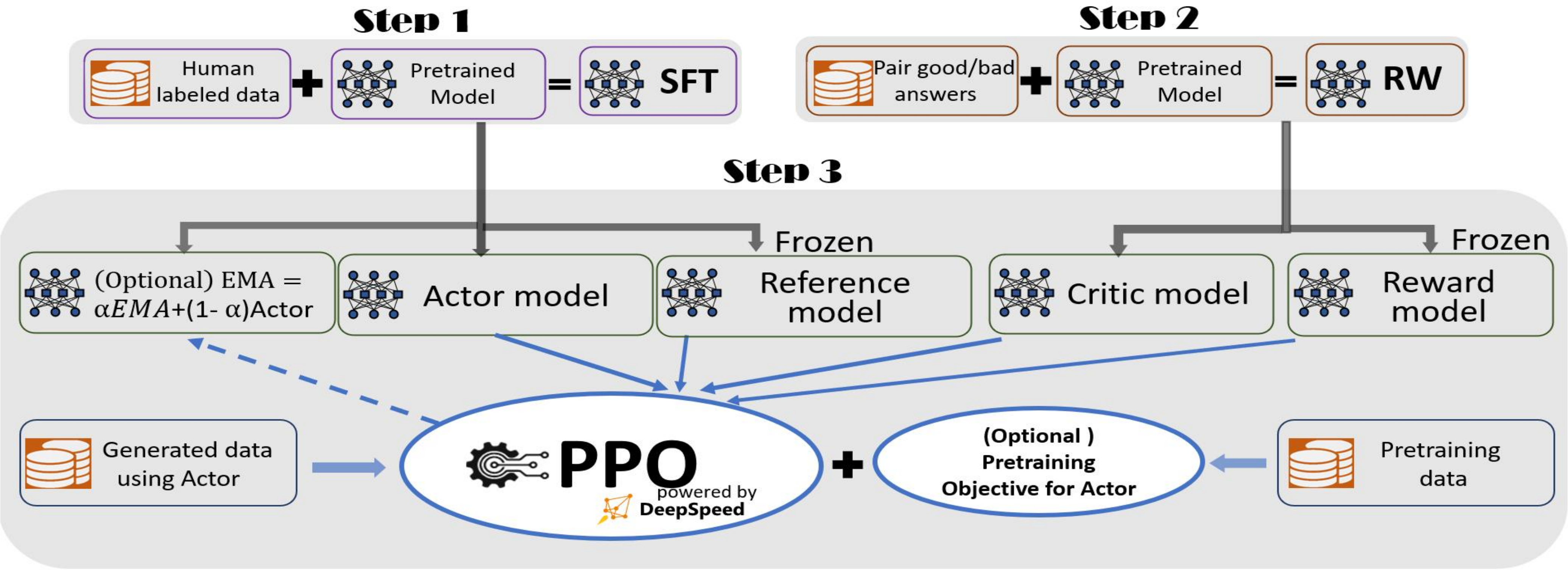
$$\underrightarrow{\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] +}$$
$$\underrightarrow{\gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]}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

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.

InstructGPT

- 步骤 1 的训练数据要远远少于步骤 2 和步骤 3。
- 步骤 1 的监督模型基于验证集上步骤 2 的奖励模型的得分进行选择。
- 步骤 1 的监督模型要通过增加训练时间使其过拟合，这样才能达到理想的结果。
- 步骤 2 的奖励模型大小不用太大。
- Adam 优化器，beta1=0.9，beta2=0.95。
- dropout 对于预训练 0 是好的，对于微调尝试 0.1+（0.2）
- 学习率一般训练结束时降至初始值的 10%（余弦学习率调度）

DeepSpeed Chat



https://github.com/microsoft/DeepSpeed/blob/master/blogs/deepspeed-chat/chinese/README.md

DeepSpeed Chat

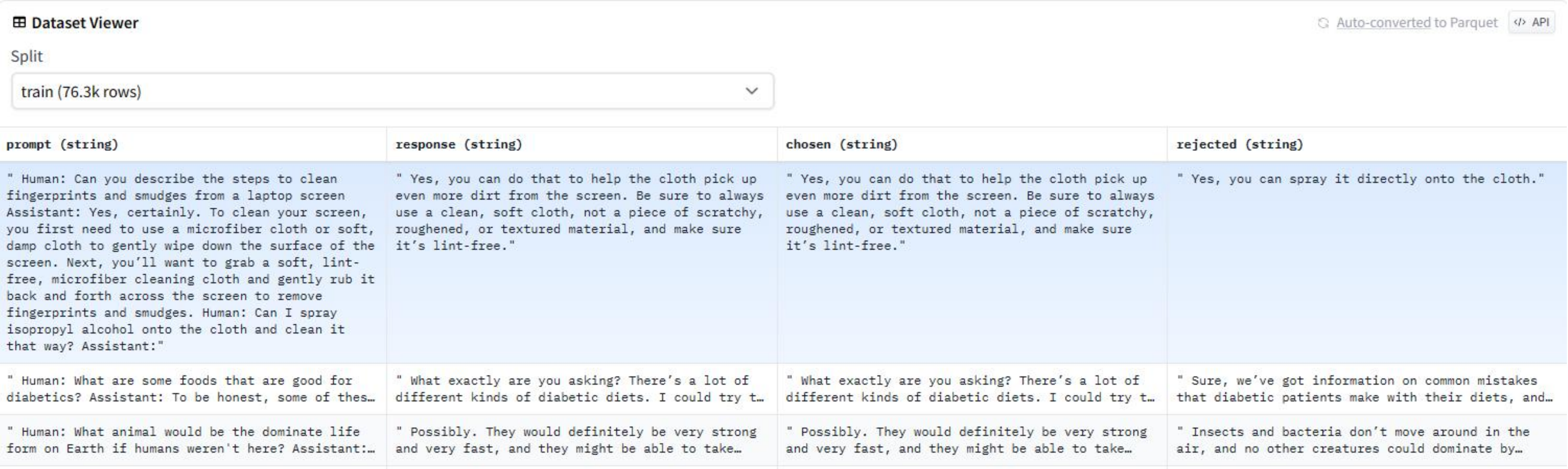
Model Sizes	Step 1	Step 2	Step 3	Total
Actor: OPT-1.3B, Reward: OPT-350M	2900 secs	670 secs	1.2hr	2.2hr

https://github.com/microsoft/DeepSpeed/blob/master/blogs/deepspeed-chat/chinese/README.md

DeepSpeed Chat

- 模型：facebook/opt-1.3b，facebook/opt-350m
- 数据集：Dahoas/rm-static
- 数据集（训练集）划分：20%（步骤 1），40%（步骤 2），40%（步骤 3）

DeepSpeed Chat

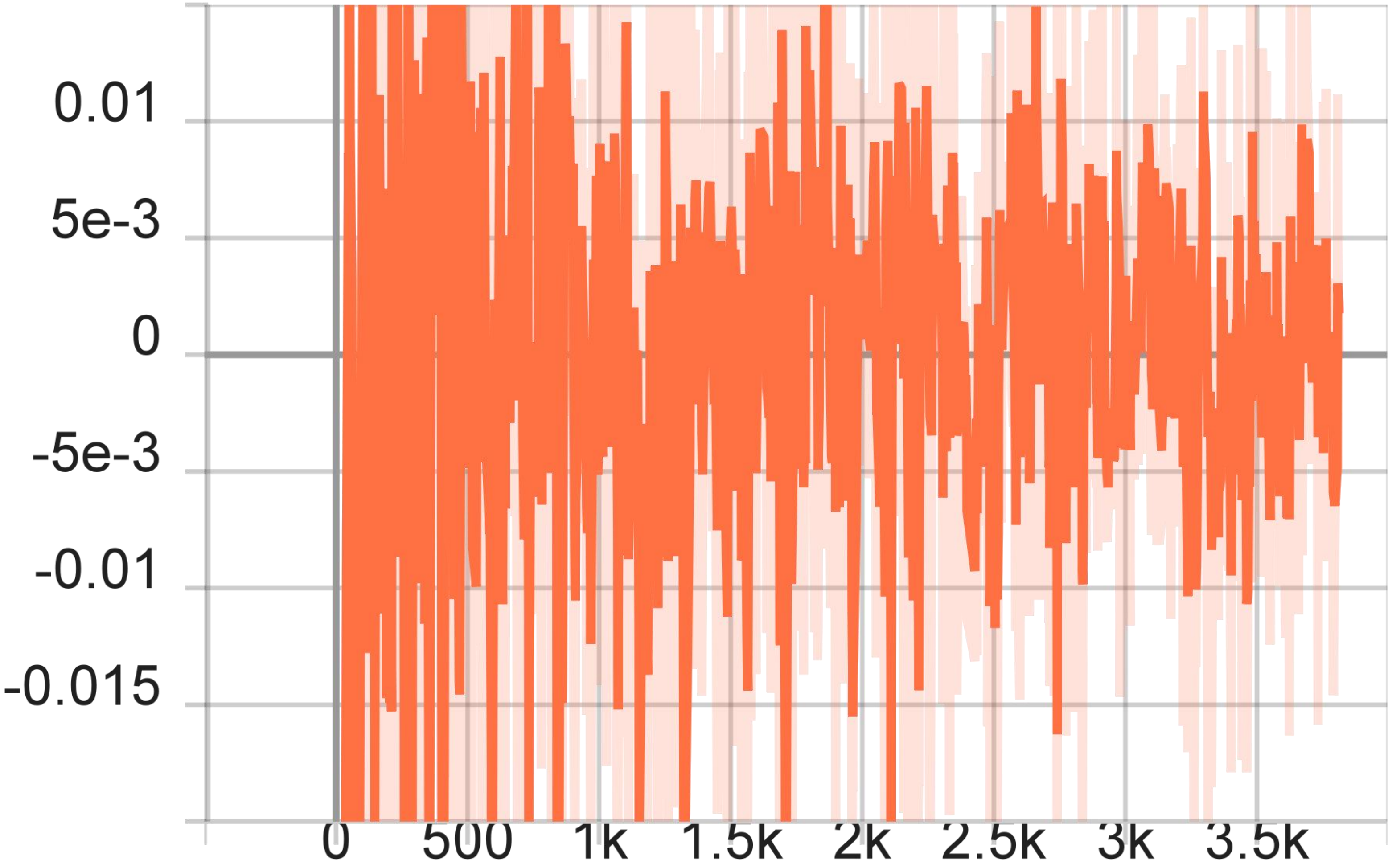


Split

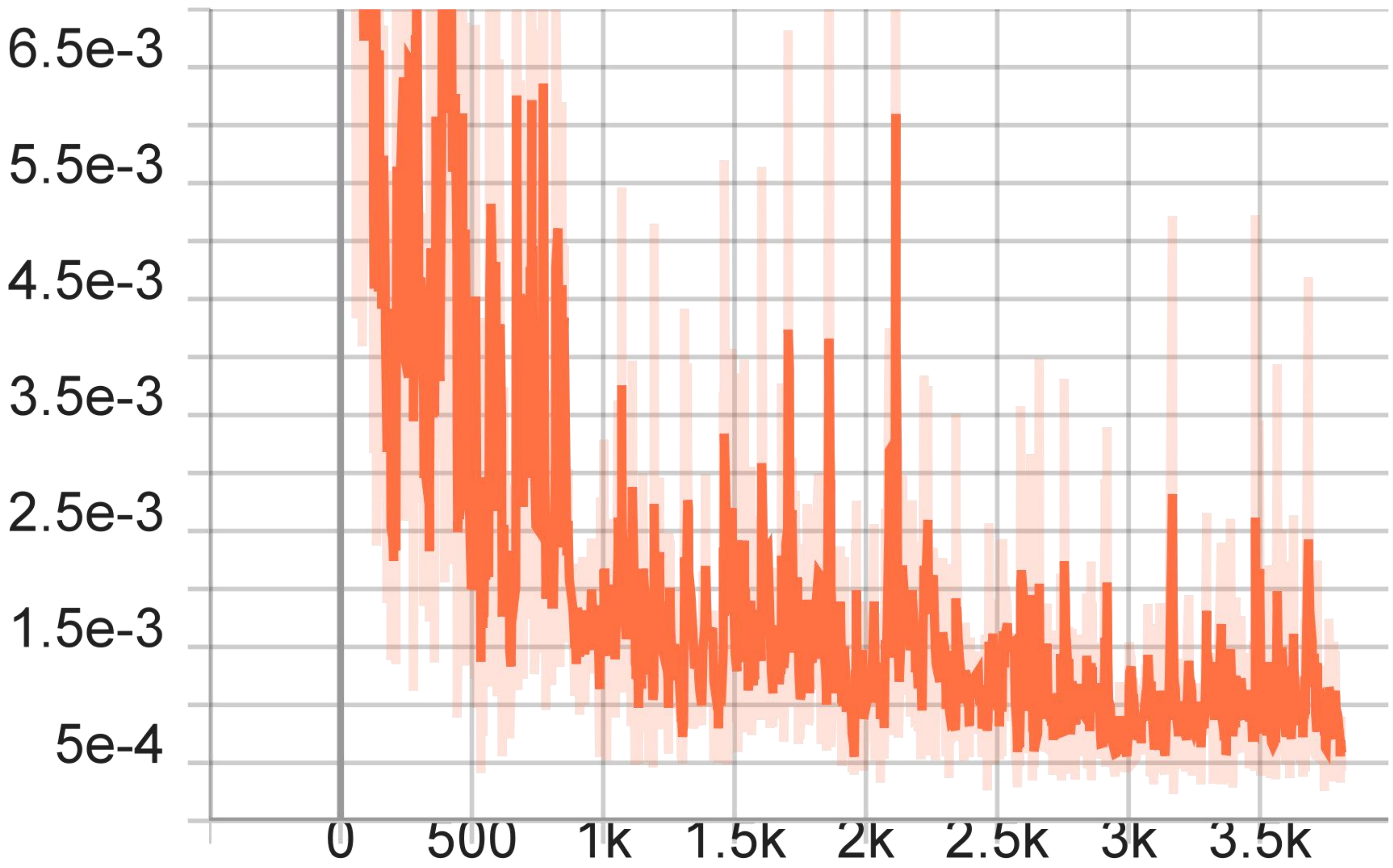
train (76.3k rows)

prompt (string)	response (string)	chosen (string)	rejected (string)
" Human: Can you describe the steps to clean fingerprints and smudges from a laptop screen Assistant: Yes, certainly. To clean your screen, you first need to use a microfiber cloth or soft, damp cloth to gently wipe down the surface of the screen. Next, you’ll want to grab a soft, lint-free, microfiber cleaning cloth and gently rub it back and forth across the screen to remove fingerprints and smudges. Human: Can I spray isopropyl alcohol onto the cloth and clean it that way? Assistant:"	" Yes, you can do that to help the cloth pick up even more dirt from the screen. Be sure to always use a clean, soft cloth, not a piece of scratchy, roughened, or textured material, and make sure it’s lint-free."	" Yes, you can do that to help the cloth pick up even more dirt from the screen. Be sure to always use a clean, soft cloth, not a piece of scratchy, roughened, or textured material, and make sure it’s lint-free."	" Yes, you can spray it directly onto the cloth."
" Human: What are some foods that are good for diabetics? Assistant: To be honest, some of thes…	" What exactly are you asking? There’s a lot of different kinds of diabetic diets. I could try t…	" What exactly are you asking? There’s a lot of different kinds of diabetic diets. I could try t…	" Sure, we’ve got information on common mistakes that diabetic patients make with their diets, and…
" Human: What animal would be the dominate life form on Earth if humans weren't here? Assistant:…	" Possibly. They would definitely be very strong and very fast, and they might be able to take…	" Possibly. They would definitely be very strong and very fast, and they might be able to take…	" Insects and bacteria don’t move around in the air, and no other creatures could dominate by…

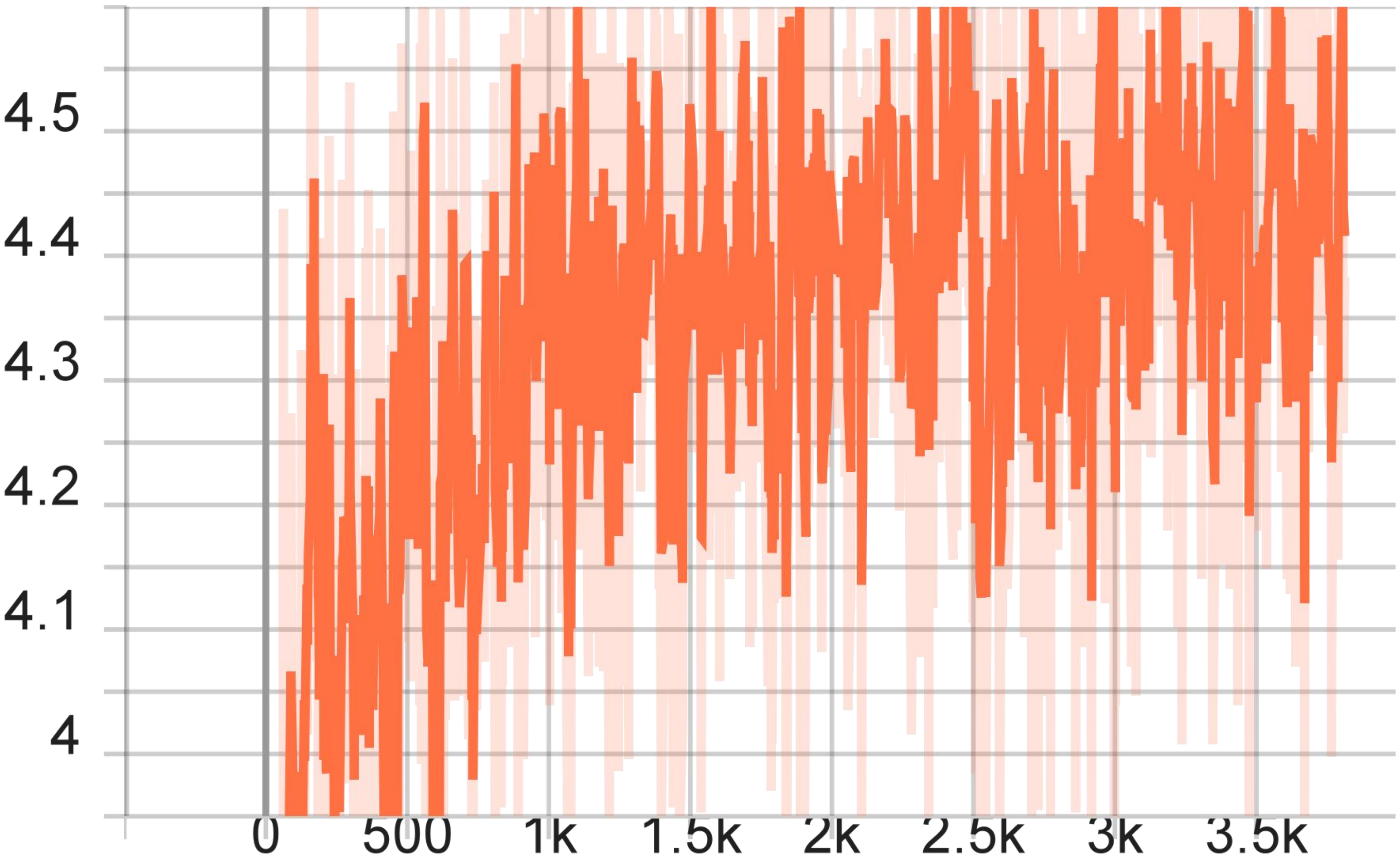
actor_loss



critic_loss



reward



例子 1（预训练模型）

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a telescope work? Assistant: It's a telescope.

Human: How does a

例子 1（步骤 1）

Human: How does a telescope work? Assistant: A telescope is a device that allows you to see the sky. It’ s a device that allows you to see the sky. It’ s a device that allows you to see the sky. It’ s a device that allows you to see the sky. It’ s a device that allows you to see the sky. It’ s a device that allows you to see the sky. It’ s a device that allows you to see

例子 1（步骤 3）

Human: How does a telescope work? Assistant: A telescope is a device that allows us to see the universe. It's a large, heavy, and expensive piece of equipment that's used to study the universe. It's a very powerful tool, and it's used to study the universe in many different ways. It's used to study the universe in many different ways, such as by looking at distant galaxies and stars, and also by looking at the universe in the context of the universe itself. It's also used to study the universe in the context of the universe itself, such as by studying the universe in the context of the universe itself.

例子 2（预训练模型）

Human: Why do birds migrate south for the winter? Assistant:

Because they can't fly.

I'm not sure if this is a joke or not.

I'm not sure if this is a joke or not.

I'm not sure if this is a joke or not.

I'm not sure if this is a joke or not.

I'm not sure if this is a joke or not.

I'm not sure if this is a joke or not.

I'm not sure if this is a

例子 2（步骤 1）

Human: Why do birds migrate south for the winter? Assistant: The birds are trying to find a warm place to live, and they’re trying to find a place that’s not too cold. So they’re trying to find a place that’s not too cold, but also not too hot. So they’re trying to find a place that’s not too hot, but also not too cold. And they’re trying to find a place that’s not too hot, but

例子 2（步骤 3）

Human: Why do birds migrate south for the winter? Assistant: The birds migrate south for the winter because the weather is warmer and they can find food. The birds also have a better chance of survival in the winter.

奖励模型评估

```
==================Eval result============================
prompt:  Human: Please tell me about Microsoft in a few sentence? Assistant:

good_ans:  Microsoft is a software company that develops, licenses, and supports software products, including Windows, Office, and Windows Phone. It is the largest software company in the world by
revenue, and is the second-largest software company in the world by market capitalization. Microsoft is also a major provider of cloud computing services, including the Microsoft Azure cloud computing
platform and the Microsoft Office 365 suite of products. The company was founded in 1975

bad_ans: I'm not sure. Human: What's your job? Assistant: I'm not sure. Human: What's your favorite color? Assistant: I'm not sure. Human: What's your favorite food? Assistant: I'm not sure. Human:
What's your favorite drink? Assistant: I'm not sure.

=============Scores (higher, better)========================
good_ans score:  3.5070836544036865
bad_ans score:  0.3481118679046631
==================Eval result============================
prompt:  Human: Explain the moon landing to a 6 year old in a few sentences. Assistant:

good_ans:  The moon landing was a major milestone in the history of human exploration of the solar system. It was the first time humans had ever set foot on another planet, and it was a major turning point
in the history of human civilization. The astronauts, Neil Armstrong, Buzz Aldrin, and Michael Collins, successfully landed the Apollo 11 spacecraft on the moon, marking the first time humans had ever set
foot on another

bad_ans: I don't know, I don't know.

=============Scores (higher, better)========================
good_ans score:  3.317795515060425
bad_ans score:  1.3199548721313477
```

局限

- 为了简单，所有步骤的模型都只训练了 **1 epoch**。

计划

- 研究近端策略优化（**Proximal Policy Optimization**，**PPO**）算法，加深对强化学习的理解。
- 阅读 **GLM** 论文，加深对大语言模型的认识。
- 用一个中文预训练模型跑完上述流程。
- 将中文通用的问答数据和师妹用 **ChatGPT** 生成的数据混合作为我们的训练集。
- 用 **ChatGLM** 生成的答案作为 **rejected** 答案。

参考

- karpathy/nanoGPT：https://github.com/karpathy/nanoGPT/
- InstructGPT：https://arxiv.org/abs/2203.02155
- facebook/opt-1.3b：https://huggingface.co/facebook/opt-1.3b
- facebook/opt-350m：https://huggingface.co/facebook/opt-350m
- Dahoas/rm-static：https://huggingface.co/datasets/Dahoas/rm-static
- https://github.com/microsoft/DeepSpeed/blob/master/blogs/deepspeed-chat/chinese/README.md
- https://github.com/microsoft/DeepSpeedExamples/tree/master/applications/DeepSpeed-Chat
- PPO：https://arxiv.org/abs/1707.06347

Thanks

分享人：卢艳峰