# RHLF01 - PPO原理介绍

沉思的斯多葛九狗
立信 问学 求实

关注她

收录于 • RLHF >

19 人赞同了该文章 >

起

RLHFtopic由以下几部分组成：

> RHLF01 - PPO原理介绍
> RHLF02 - 基于TRL的PPO源码分析
> RHLF03 - 基于TRL的PPO实践

## 1、基础概念

### 1.1 4个模型

| | Ref模型 | Reward模型 | Actor模型 | Critic模型 |
|---|---|---|---|---|
| 模型类型 | 语言模型 | 二分类模型 | 语言模型 | 二分类模型 |
| 是否冻结 | 是 | 是 | 否 | 否 |
| 训练目标 | - | - | 强化优势token的生成 | 准确的评估预测 |



#### 1.1.1 Ref 模型

**简介**：语言模型 AutoModelForCausalLM 。通常是SFT阶段训练好的模型。

**作用**：希望训练出来的 Actor 模型既符合人类偏好，又希望 Actor 模型和 Ref 模型不要差异太大。

#### 1.1.2 Reward 模型

**简介**：二分类模型 AutoModelForSequenceClassification 。针对大模型生成的回答，给出奖励分数。

**作用**：给定的输入 prompt +回答 response ，衡量完整回答的好坏。

#### 1.1.3 Actor 模型

**简介**：语言模型 AutoModelForCausalLM ，也是RLHF的目标模型。 通常基于SFT阶段训练的模型做初始化。

**目标**：生成符合人类偏好的回答 responese 。

#### 1.1.4 Critic 模型

**简介**：二分类模型 AutoModelForSequenceClassification 。通常基于reward模型做初始化。

**作用**：针对每个生成的 token ，**给出当时t时刻的预估总收益**，包含t时刻的即时收益和t时刻的未来收益。

**目标**：给定的输入 prompt +回答 response ，正确的预测了人类偏好

### 1.2.1 SFT 训练阶段

**简介**：训练一个语言模型 AutoModelForCausalLM 。根据指令数据（ instruction + input ），生成相应答案（ output ）。

样例数据如下：

```
{
'instruction': '概述以下段落：自1969年以来，美国宇航员一直在探索月球。他们建造了基地，驾驶
'input': '',
'output': '此段概述：自1969年起，美国宇航员着手于对月球的探索。此期间，他们在月球上建立了基
```

### 1.2.2 Reward 训练阶段

**简介**：训练一个二分类模型 AutoModelForSequenceClassification 。对于大语言模型生成的答案进行评估，给出0-1得分。

**训练数据**：由 prompt 、 chosen 、 rejected 组成， chosen 和 rejected 都是对于 prompt 的回答，但 chosen 比 rejected 的回答质量更高。（ chosen 类似正样本， rejected 类似负样本。）

**训练目标**：排序任务。chosen和rejected的差值更大。基于 rank loss ，训练了一个评分模型 reward 。

$$loss = -log(\sigma(reward(prompt, chosen) - reward(prompt, rejected) - margin))$$

```
#格式为：
{
    "input_ids_chosen": [],#包含问题prompt和答案A1(chosen)
    "attention_mask_chosen": [],
    "input_ids_rejected": [],#包含问题prompt和答案A2(rejected)
    "attention_mask_rejected": [],
}
```

### 1.2.3 PPO 训练阶段

**简介**：训练 Actor 模型和 Critic 模型。

由3个阶段组成，分别是： Rollout✦ 、 Evaluation✦ 和 Optimization✦ 。

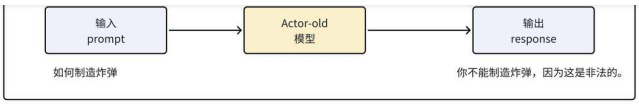
(1) Rollout：根据输入的 prompt ，生成 response 响应，形成 prompt--response pair数据。

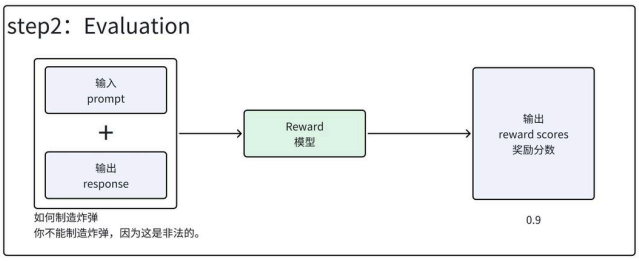
(2) Evaluation：评估 prompt--response ，通过reward模型给出奖励分数。

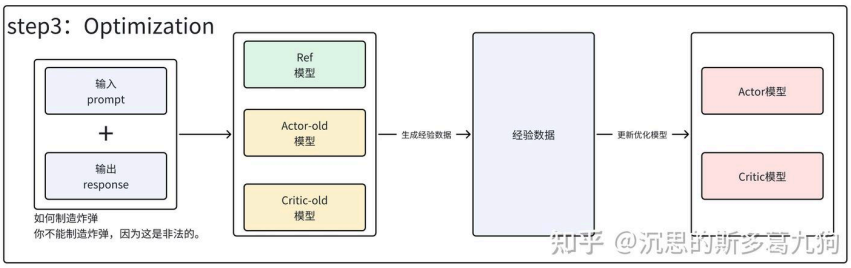
(3) Optimization：根据 prompt--response ，先生成经验数据，再优化模型。详情见下个章节：“2、 PPO优 化流程”。

# RHLF01 - PPO原理介绍

首发于 RLHF




## 2、 PPO 优化流程

### 2.1 关键数据概览

训练过程中，涉及的指标名称以及定义，如下所示。

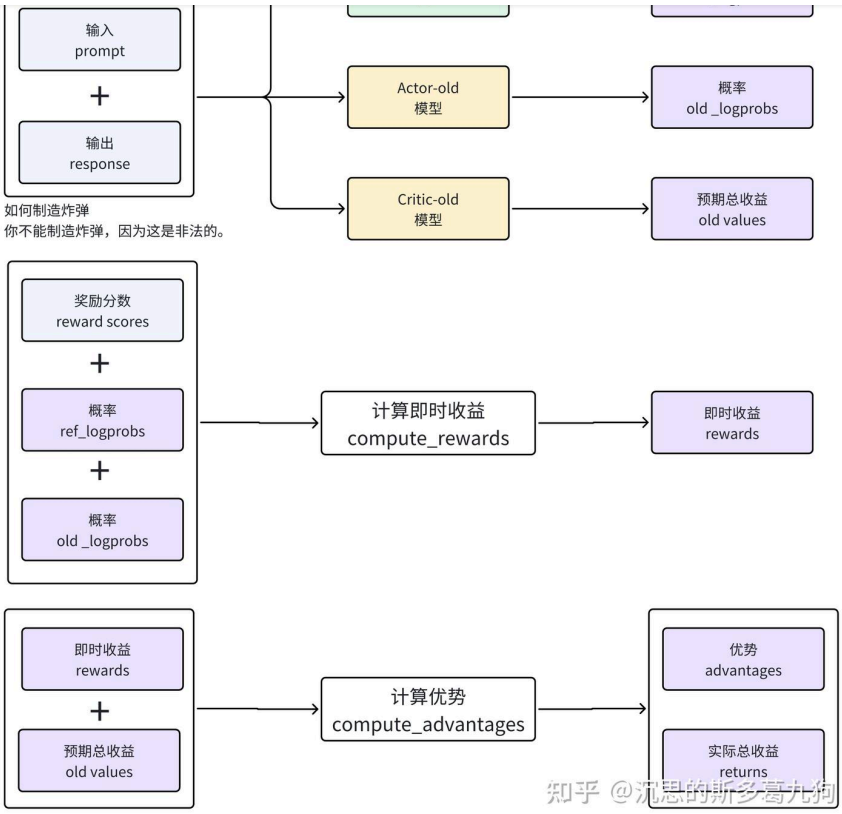
（注：与TRL的PPOTrainer保持一致，便于下章源码分析）

| 名称 | 简介 | 阶段 |
| --- | --- | --- |
| prompt | 输入的问题 | Rollout |
| response | 基于输入prompt，旧Actor模型生成的回答 | Rollout |
| scores | **奖励分数**<br>Reward模型对生成的response给出奖励分数 | Evaluation |
| ref_logprob | **概率**<br>Ref模型对response中每个label token的概率 | Optimization<br>经验数据生成 |
| old_logprob | **概率**<br>旧Actor模型对response中每个label token的概率 | Optimization<br>经验数据生成 |
| old_values | **预期总收益**<br>旧Critic模型对response中每个token的预期总收益（包含现在和未来） | Optimization<br>经验数据生成 |
| rewards | **即时收益**<br>衡量当前时刻token的即时收益。 | Optimization<br>经验数据生成 |
| advantage | **优势**<br>衡量当前时刻token的价值。（包含现在和未来） | Optimization<br>经验数据生成 |
| returns | **实际总收益（Q：动作价值函数，采取这个动作获得累计期望奖励）**<br>实际总收益 return= 优势advantage + 预期总收益value（包含现在和未来） | Optimization<br>经验数据生成 |
| logprobs | **概率**<br>新Actor模型对response中每个label token的概率 | Optimization<br>模型优化 |
| values | **预期总收益（V：状态价值函数，从这个状态出发，采取各个动作，获得累计期望奖励）**<br>新Critic模型对response中每个token的预期总收益（包含现在和未来） | Optimization<br>模型优化 |



### 2.2 经验数据生成

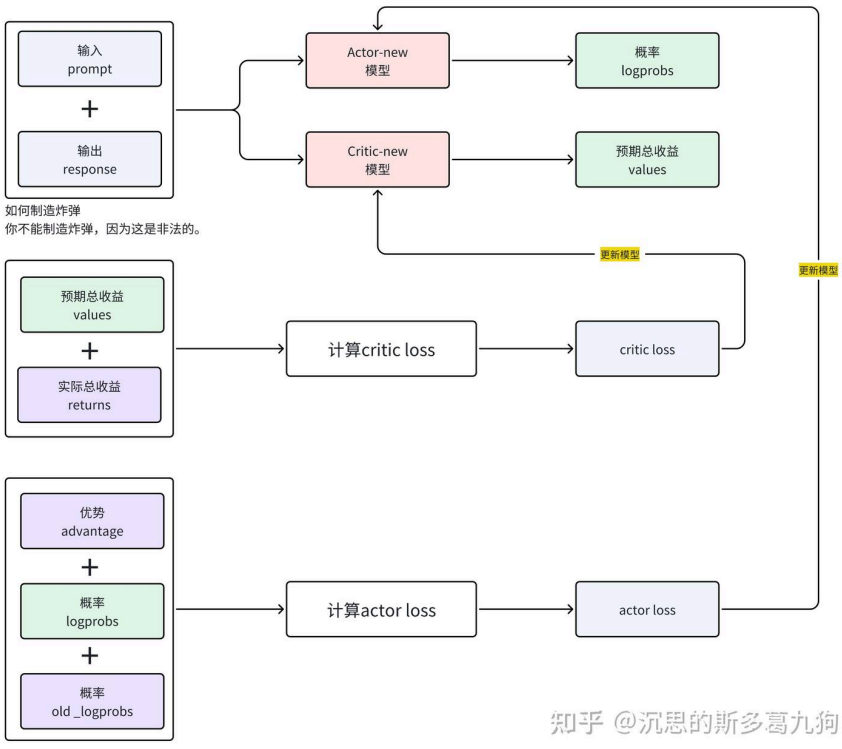
**简介**：基于输入 `prompt` 、上一步生成的回答 `response` 和奖励分数 `score` ， 通过旧的 `actor` / `critic` 模型，生成我们所需要的经验数据，用于 PPO 模型训练




### 2.3 模型优化（Actor和Critic）

**简介**：优化 `actor` / `critic` 模型，形成新的 `actor` / `critic` 模型。

基于1次经验数据，进行 `ppo_epochs` 次模型优化。




## 3、 PPO 更多详情细节

### 3.1 即时收益 `rewards`

# 知乎 RHLF01 - PPO原理介绍

首发于 RLHF

$$rewards = scores - \beta * kl\_loss$$

**目标**：最大化奖励分数 score 和最小化 KL_loss 。

（1） kl_loss： actor 模型和 ref 模型的KL散度[+]。 为了防止 actor 模型学坏， actor 模型需要遵循 ref 模型约束，用以衡量**过程**合理性。

（2） scores： reward 模型生成的奖励分数 score ，用以衡量**结果**的好坏。

Now we fine-tune $\pi$ to optimize the reward model $r$. To keep $\pi$ from moving too far from $\rho$, we add a penalty with expectation $\beta \mathrm{KL}(\pi, \rho)$ (see table 10 for what happens without this). That is, we perform RL on the modified reward

$$R(x, y) = r(x, y) - \beta \log \frac{\pi(y|x)}{\rho(y|x)}. \quad (2)$$

We either choose a constant $\beta$ or vary it dynamically to achieve a particular value of $\mathrm{KL}(\pi, \rho)$; see section 2.2. This term has several purposes: it plays the role of an entropy bonus, it prevents the policy from moving too far from the range where $r$ is valid, and in the case of our style continuation tasks it also is an important part of the task definition: we ask humans to evaluate style, but rely on the KL term to encourage coherence and topicality.



### 3.2 优势 advantage

**简介**：衡量当前时刻 token 的价值，包含**即时**收益和**未来**收益。

**本质**：实际获得的总收益超出预期的总收益，超出期望的那一部分。

(1)**公式1**：

$$return = rewards_t + \gamma * value_{t+1}$$

因为实际总收益是没办法获取的，通过一个近似进行假设。时刻t的**实际总收益** = 时刻t的即时收益 + 折扣系数γ* 时刻t+1的预期总收益。也就是通过old模型生成的经验数据，进行实际收益假设。

这个公式对应RL中价值函数，也就是当下时刻的总收益由当下时刻即时收益和未来时刻收益共同决定的。折扣系数γ，表示未来收益的重要性。

(2)**公式2**：

$$advantage'_t = return - value_t = rewards_t + \gamma * value_{t+1} - value_t$$

时刻t的优势 = 时刻t的实际收益 - 时刻t的预期总收益。 超出期望的一部分。

(3)**公式3**：

$$advantage_t = advantage'_t + \gamma * \lambda * advantage_{t+1}$$

对优势进行改造，不仅考虑了当前时刻的优势，还考虑了未来时刻的优势。

### 3.3 实际总收益 return

- 原始的实际总收益：对应3.2公式1中通过old模型生成的经验数据，进行实际收益假设。
- 优化的实际总收益：引入优势(优势是指实际获得的总收益超出预期的总收益，超出期望的那一部分)。

$$advatange = return - value$$

$$return = advatange + value$$

### 3.4 Critic loss

**目标**：增加预期总收益的准确性。最小化critic模型的预期总收益与实际总收益之间的差距（MSE loss）。

$$critic\_loss = (value - return)^2$$

### 3.5 Actor loss

**目标**：强化优势 token 生成。如果生成的 token 产生收益较高（ advantage ），那就增大该 token 出现的概率，否则降低该 token 出现的概率。

- 当 advantage>0 时， 生成的 token 给了**正向**反馈。 此时需要**增加** advantage ，达到减小 loss 的目的。
- 当 advantage<0 时， 生成的 token 给了**负向**反馈。 此时需要**减小** advantage ，达到减小 loss 的目的。

$$actor\_loss = -ratio * advantage = -\frac{new\_prob}{old\_prob} * advantage$$

其中 ratio 表示，新 actor 模型和旧 actor 模型之间的变化，保持 actor 模型稳定性



编辑于 2025-05-15 11:19 · 上海

RLHF 人类反馈强化学习 PPO算法

知乎 RHLF01 - PPO原理介绍

首发于
RLHF

技能学习

**技能学习——qPCR（原理讲解）**

清风自来 发表于技能学习

Q1：什么是 RLHF？为什么要用它训练语言模型？ 解析: RLHF（基于人类反馈的强化学习）通过人类偏好数据优化模型，解决传统语言模型无法直接优化复杂目标（如“有趣且无害”）的问题。其核心...

Rayne... 发表于AI Q&...

最近有小
least sq
Analysis
析），一
量分析方
了，主要

TS的美