

1. 训练数据

1.1 训练数据集

基于以下开源数据集构建的 GRPO 训练样本:

https://huggingface.co/datasets/notbadai/math_reasoning

1.2 构建训练数据

1.2.1 第一种：按 **GSM8K** 的格式处理

在原始 question 后面添加以下指令，并要求最终结果在 “####” 后输出，指令如下：

```
instruction_following = 'Let\'s think step by step and output the final answer after  
"####".'
```

具体代码如下：

https://github.com/jouw/verl/blob/train-openseek/examples/data_preprocess/build_notbadai_math_reasoning.py

但发现 verl 在抽取回答计算 score 时，只抽取数值类型数据，而训练数据中有非数值类型的回答，则不能处理。因此改换成接下来的第二种方式。

训练时处理 gsm8k 回答的代码如下：

- 计算 score:

https://github.com/jouw/verl/blob/train-openseek/verl/utils/reward_score/__init__.py
中第 43-46 行。

- 抽取回答:

https://github.com/jouw/verl/blob/train-openseek/verl/utils/reward_score/__init__.py

1.2.2 第二种（改进后的方式）：按 \boxed{} 格式处理回答

在原始 question 后添加以下指令：

```
instruction_following = "Let's think step by step and output the final answer within\n\\boxed{}."
```

具体代码如下：

https://github.com/jouw/verl/blob/train-openseek/examples/data_preprocess/build_notbadai_math_reasoning_refine.py

2. 修改 verl 的模型训练源码

基于 verl 源码 fork 新分支后，修改的训练代码已经上传到 github，代码仓库是：<https://github.com/jouw/verl/tree/train-openseek>，具体修改在分支 train-openseek 中。

修改的部分主要是模型保存和 tokenizer，这两部分的逻辑跟模型不兼容，修改的具体差异对比如下：

<https://github.com/volcengine/verl/compare/main...jouw:verl:train-openseek>

3. 模型训练

3.1 训练第一版模型

3.1.1 训练数据

使用的是章节 1.2.1 中按 GSM8K 方式构建的训练样本

3.1.2 训练脚本

训练脚本已经上传 github 如下:

<https://github.com/jouw/OpenSeek/blob/competition-jouw/openseek/competition/pz/jouw/final-round/train-grpo.sh>

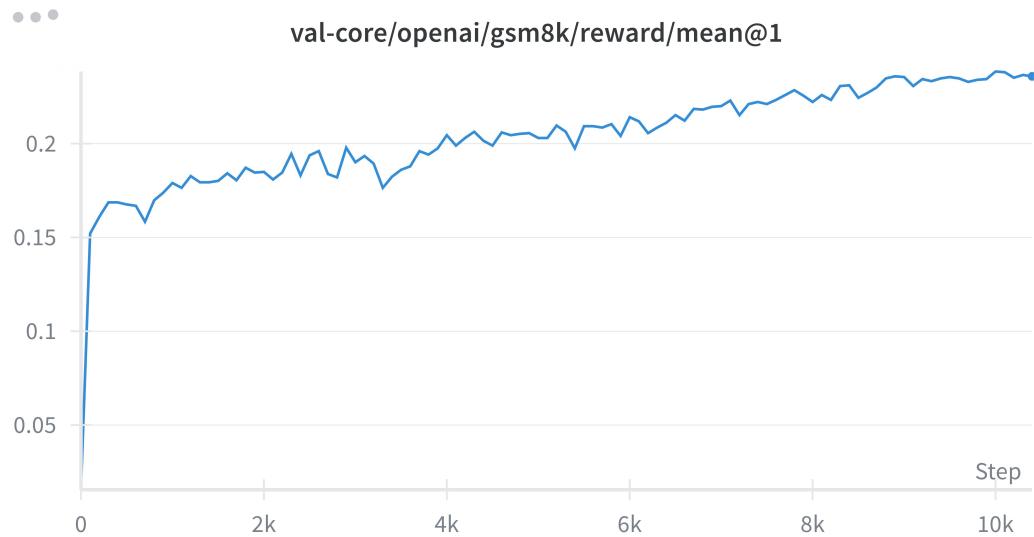
说明:

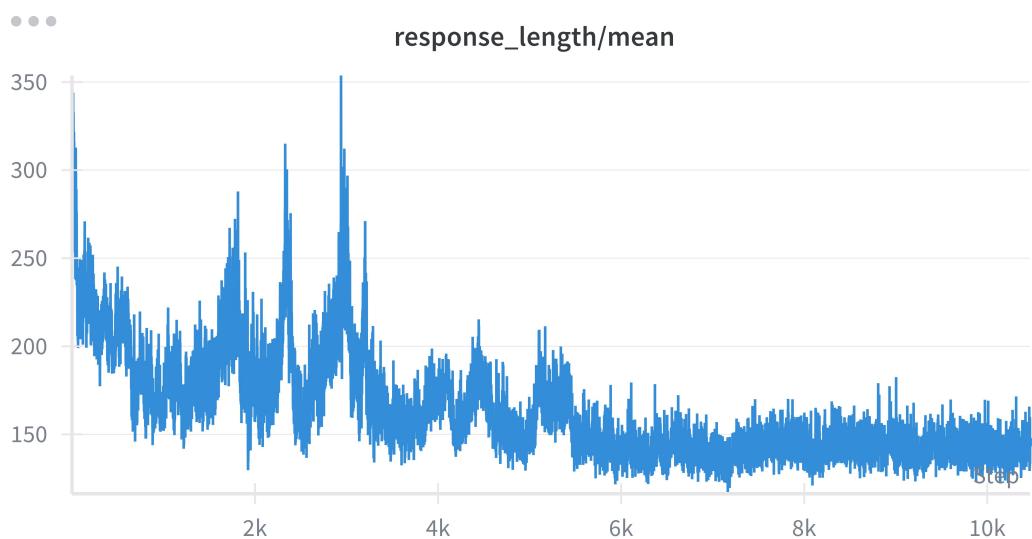
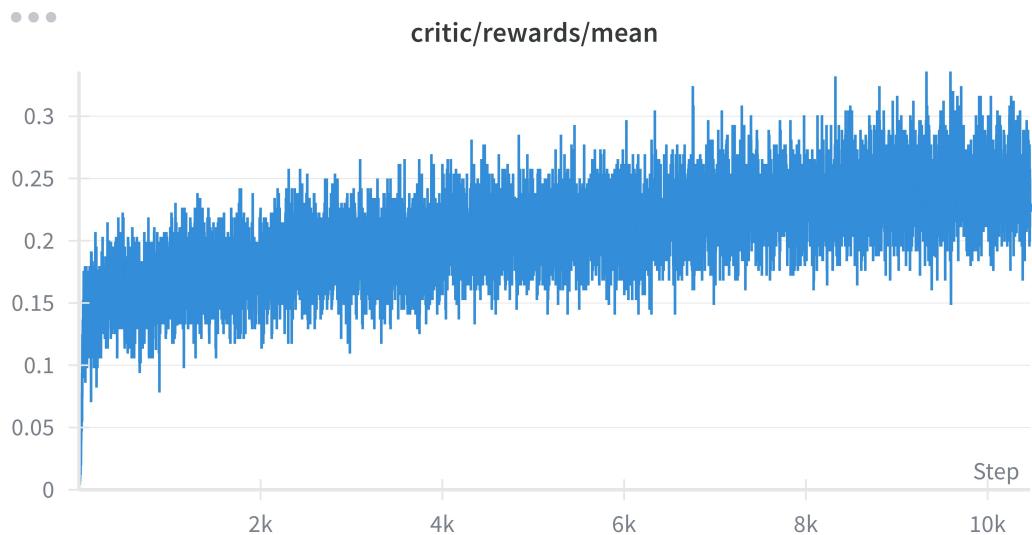
(1) 使用 GRPO 的训练策略

```
algorithm.adv_estimator=grpo \
```

(2) actor 模型的 lr 使用 cosine 衰减

```
actor_rollout_ref.actor.optim.lr_warmup_steps=20 \
actor_rollout_ref.actor.optim.warmup_style="cosine" \
actor_rollout_ref.actor.optim.min_lr_ratio=0.1 \
actor_rollout_ref.actor.optim.num_cycles=0.5 \
actor_rollout_ref.actor.optim.weight_decay=0.01 \
```





wandb 训练日志:

https://wandb.ai/oceanplus/OpenSeek-Small-v1-SFT_RL_Training/reports/The-GRPO-training--VmIldzoxNDM2NDUwMw

3.2 训练改进版模型

3.2.1 训练数据

采用章节 1.2.2 中的训练数据处理方式，对于数值和非数值的回答都能抽取。

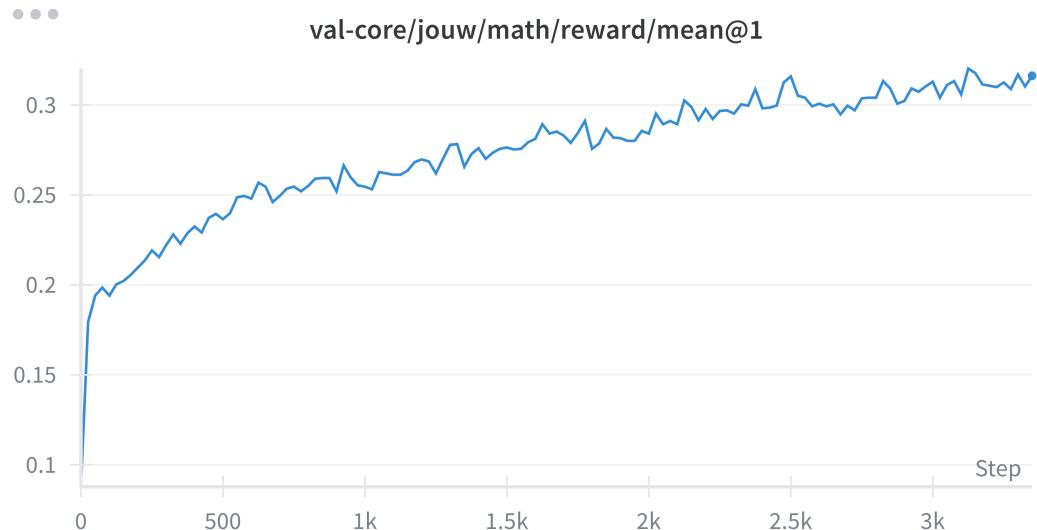
3.2.2 训练脚本

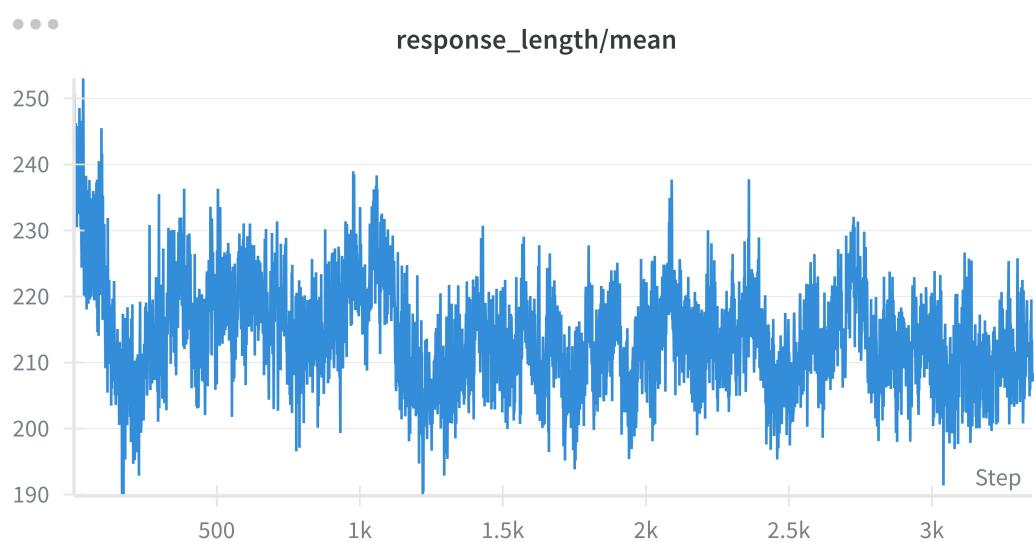
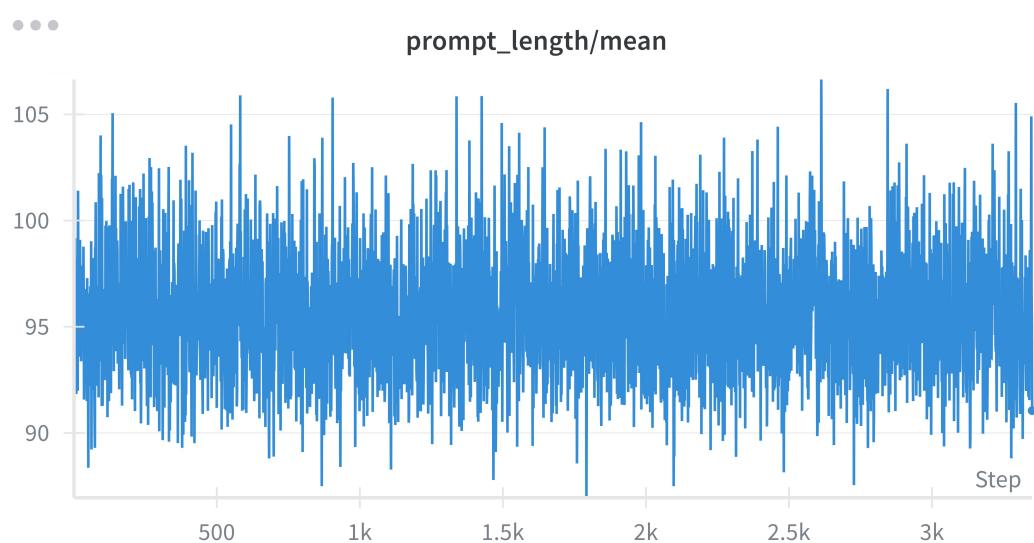
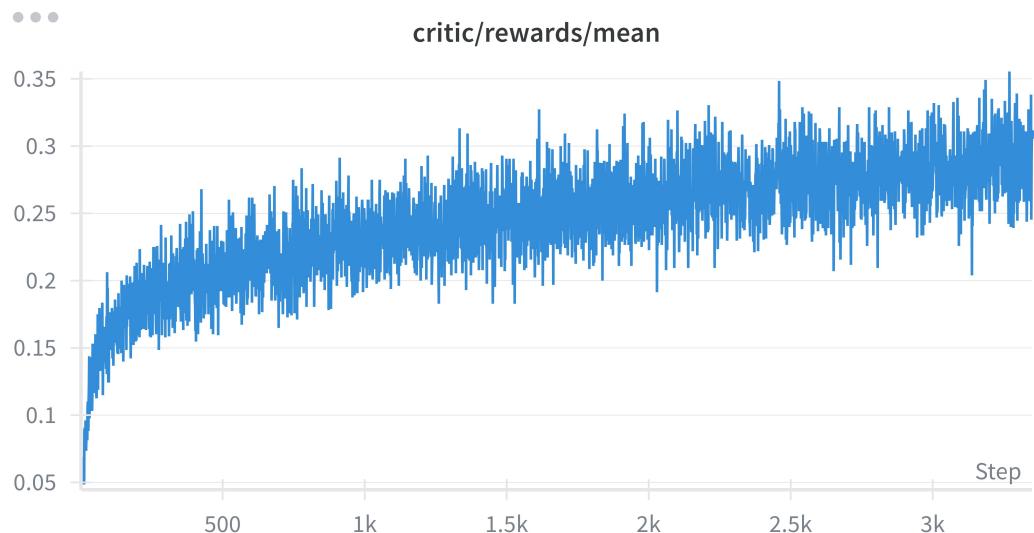
训练脚本已经上传 [github](https://github.com/jouw/OpenSeek/blob/competition-jouw/openseek/competition/pz/jouw/final-round/train-grpo-v4.sh) 如下:

<https://github.com/jouw/OpenSeek/blob/competition-jouw/openseek/competition/pz/jouw/final-round/train-grpo-v4.sh>

相比第一版主要做了以下调整:

- (1) 数据集换用 `\boxed{}` 的方式, 对数值和非数值都能处理
- (2) 配置 `actor_rollout_ref.rollout.n=5`



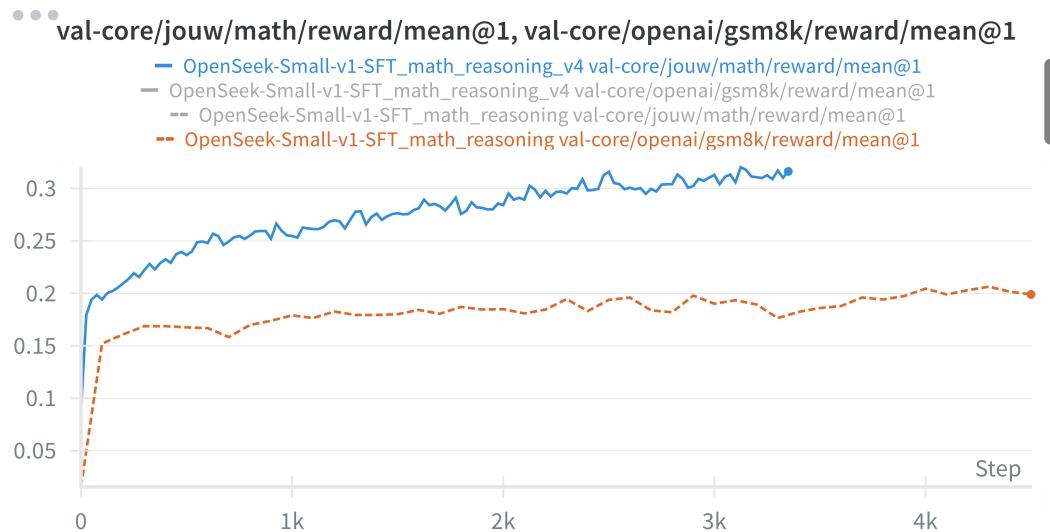


wandb 训练日志:

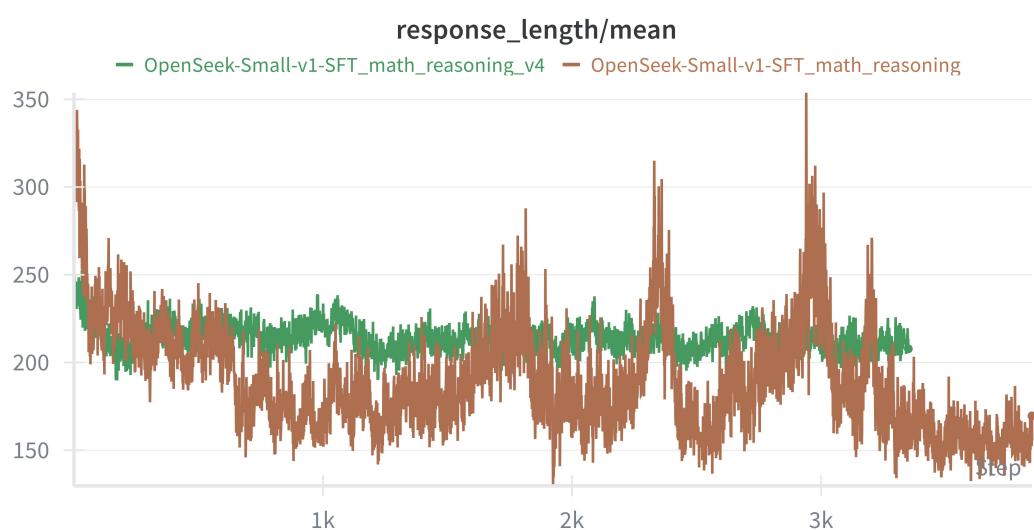
https://wandb.ai/oceanplus/OpenSeek-Small-v1-SFT_RL_Training/reports/The-GRPO-training-v4--VmldzoxNDM2NDE0Mg

3.3 两版模型训练对比

(1) 橙色线是第一版模型，蓝色线是改进版模型，可以看出改进后的版本的准确率明显更高



(2) 返回长度上，改进版模型 (math_reasoning_v4) 比第一版 (math_reasoning) 整体上要长，说明它的推理思考更长。



4. 效果评测

数据集	第一版模型	改进版模型
minerva_math	4.77941	5.51471
gsm8k	26.23199	33.28279
amc23	20	12.5
olympiadbench	4.74074	5.03704
math500	15.6	23.8
aime24	0	3.33333
平均分	11.89202333	13.911