上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

cs2916-2025 第一次大作业实验报告

学院 致远学院

班级 电院 2331ACM

学号 523030910148

姓名 庄裕旻

2025 年 4 月 27 日

摘 要

本次大作业基于 Qwen2.5-Math-1.5B 模型进行实验，主要包括 SFT 和 GRPO 两部分的内容。在实验中，我们补全了 GRPO 部分的 reward 计算和 loss 计算，并修改了 SFT 和 evaluation 的代码，使得可以高效地实验不同超参数对 short cot sft 的影响。实验结果表明，我们的方法在多个数据集上取得了优于 baseline 的表现，long cot、grpo 严格优于 baseline，short cot 和 baseline 相当。

目录

1 SFT

本节介绍基于 Qwen2.5-Math-1.5B 模型的 SFT 实验，包括 short cot 和 long cot 两种方式。

1.1 Short COT

在 short cot 实验中，我们修改了原有代码，使其能够高效地实验不同超参数对训练效果的影响。实验结果显示，部分指标超过了 short cot sft baseline。

由于 Short CoT 的参数调节较为复杂，因此需要对可能的状态空间进行搜索

首先进行粗略搜索，状态空间为：

```
evaluation_epochs = [3, 6]
explore_batch_size = [128, 256, 512]
explore_lr = [1e-5, 2e-5, 4e-5]
```

按照 4 个指标相对 baseline 比值的调和平均数（会放大最小值的影响）作为评价分数得到的结果如图 1所示。第一轮得到的最大值: 1.0758890919394155, 对应的参数: (512,

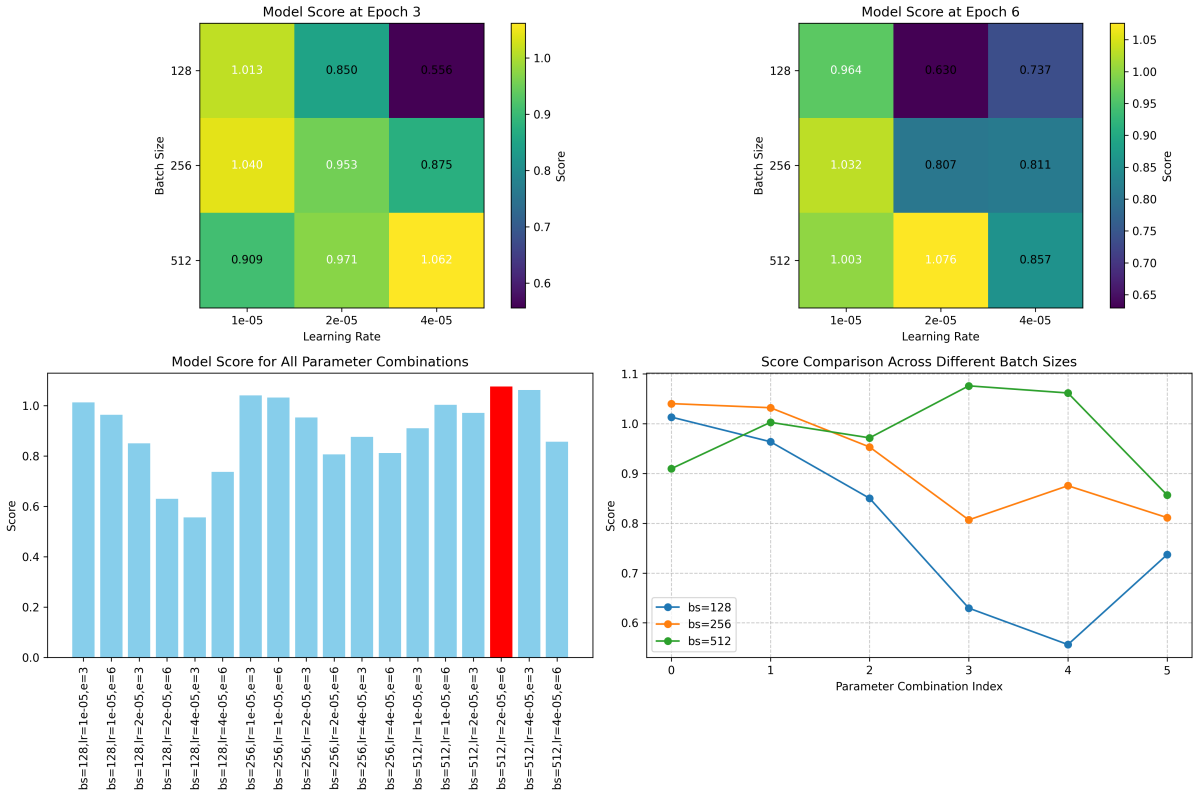

图 1: short cot sft 超参数第一轮搜索

2e-05, 6)，具体表现如下：

- AMC23 acc: 0.4

- GSM8k acc: 0.707
- MATH500 acc: 0.412
- OlympiadBench acc: 0.108

其中有 2 个指标严格大于 baseline，1 个指标略小于 baseline，1 个指标小于 baseline，考虑到正确率比值的调和平均值 $1.07 > 1$，可以认为这个模型的能力和 baseline 近似相当。

1.2 Long COT

在 long cot 实验中，我们的模型在所有指标上全部超过了 long cot sft baseline，具体表现如下：

- AMC23 acc: 0.325
- GSM8k acc: 0.694
- MATH500 acc: 0.408
- OlympiadBench acc: 0.138

训练参数设置为：BS=128, EP=6, LR=2e-5。

2 GRPO

本节介绍基于 Qwen2.5-Math-1.5B 模型的 GRPO 实验。我们的方法在所有指标上全部超过了 grpo baseline，具体表现如下：

- AMC23 acc: 0.525
- GSM8k acc: 0.775
- MATH500 acc: 0.618
- OlympiadBench acc: 0.246

2.1 训练流程

整个训练的流程是在 GRPO 的基础上进行人工划分阶段的 curriculum learning，主要分为基本训练和质量提升训练两个阶段。超参数是直接使用框架里给的超参数。

2.1.1 相关探索

- 修改题目的难度配比：几乎没有效果，甚至有负面影响
- 从 sft 的结果开始训练：几乎没有效果
- 使用 DAPO 的论文 [1] 里提到的 Clip-Higher：有效果

2.1.2 基本训练

在基本训练阶段，我们使用默认数据集和默认 LOSS 函数，设计了如下 reward 函数：

- 回答无法解析出答案：-1
- 答案错误：0
- 答案正确：100

基本训练过程分为 3 轮，每一轮结束后，人工通过 reward 等指标的变化判断“最有潜力”的 checkpoint 最为下一轮的训练起点。

2.1.3 质量提升训练

在质量提升训练阶段，我们使用默认数据集和默认 LOSS 函数，设计了更为复杂的 reward 函数：

- 回答无法解析出答案：-10
- 答案错误：0
- 答案正确：100
- 思维链长度 bonus：在回答中没有重复的情况下，提供一个不超过 10 的长度 bonus
- 语言 bonus：不出现中英混杂的情况时，可获得 5 分的 bonus

质量提升训练共 2 轮，其中第 2 轮打开了 Clip-Higher。

A 如何复现与具体代码、数据

仓库地址：https://github.com/happyZYM/cs2916-2025，实验日志：https://wandb.ai/zymx/cs2916-2025

- shortcot: 运行 *scripts/explore.py*，或者直接用对应超参数训练。效果可能有一定差别。
- longcot: 运行 *scripts/sft_longcot.sh*。效果可能有一定差别。
- grpo：无法一键复现，请根据前文讲述的过程调节 *scripts/grpo.sh*、reward、loss 相关代码。同时，由于 RL 的随机性相当大，不保证效果能完全达到前文所述的效果，验证请从后文的链接下载 checkpoint 进行验证。

short COT、long cot、grpo 三部分训练出的模型 checkpoint 可以通过以下方式下载：

- 交大云盘：链接https://pan.sjtu.edu.cn/web/share/860110312f97f0a9673f0f2dc050644c，提取码: uqrz
- 备用下载方式：
  - shortcot sft：https://alist-cf.zhuangyumin.dev/x/share/cs2916-2025-hw1/sft_shortcot_512_2e-5_6.tar.zst?sign=DVfROmggJOrchjSJJhM8HgnK_9cK8LkXgXX_ASMkny4=:0
  - longcot sft：https://alist-cf.zhuangyumin.dev/x/share/cs2916-2025-hw1/sft_longcot5.tar.zst?sign=JQmlhJpzXMF87NaLvVEluwUwYGX7vuHJsMzcfryUnyY=:0
  - grpo: https://alist-cf.zhuangyumin.dev/x/share/cs2916-2025-hw1/grpo_quality_round2.tar.zst?sign=4ftP-CJSHMHLSlCZkD8cj4ZltxQBZvkNxSjEhHXjCxo=:0

B REFERENCES

[1] Q. Yu et al. “DAPO: An Open-Source LLM Reinforcement Learning System at Scale.” arXiv: 2503.14476 [cs], Accessed: Apr. 5, 2025. [Online]. Available: http://arxiv.org/abs/2503.14476, pre-published (cit. on p. 5).

[2] X. Li, H. Zou, and P. Liu. “LIMR: Less is More for RL Scaling.” arXiv: 2502.11886 [cs], Accessed: Apr. 10, 2025. [Online]. Available: http://arxiv.org/abs/2502.11886, pre-published.