OpenSeek 共创大模型挑战赛决赛技术报告

1. 数据工程

1.1 数据选取与验证

数据集：本次实验的核心数据源为 Big-Math-RL-Verified-Processed 数据集，具体链接为：https://www.modelscope.cn/datasets/open-r1/Big-Math-RL-Verified-Processed，共包含 215,608 条样本。

数据污染核查： 秉持严谨的学术精神，在撰写报告时我对训练集与官方测试集进行了精确匹配核查，具体如表 1。核查发现，二者存在 30 道题目的重合，主要分布在 aime24 (13 道) 和 amc23 (17 道) 两个测试集中，总体来讲占比约 1%左右。

表 1 数据泄露核查

测试集名称	重复数量
aime24	13
amc23	17
math500	0
minerva_math	0
gsm8k	0
olympiadbench	0

1.2 数据格式构建

为适配强化学习框架，我参考 OpenSeek-SFT 的 prompt 结构，构建了标准化的数据格式。其中，prompt_content 字段通过指令引导模型进行分步推理，并要求将最终答案置于 \boxed{} 中，这为后续奖励函数的设计奠定了基础。具体为：

```
{
    "data_source": 'xxxxxxxx'
    "prompt": [
        {
            "role": "user",
            "content": prompt_content,
        }
    ],
    "ability": "math",
    "reward_model": reward_model_data,
    "extra_info": {
        "index": idx,
        "original_problem": problem_raw,
        "domain": domain,
        "llama8b_solve_rate": solve_rate,
    },
}
```

其中：prompt_content = instruction+ " " + problem_raw ， instruction = r'Please reason step by step,and must put your final answer within \boxed{}.Question:'

2. 模型训练方法

2.1 奖励函数设计 (Reward Function)

我的奖励函数旨在引导模型同时优化解题的规范性和准确性，其构成如下：

格式奖励 (Format Reward, 权重 10%)： 检查模型的输出是否包含 \boxed{} 结构。若匹配，则奖励为 1，否则为 0。此项旨在激励模型学会按要求格式作答。

正确性奖励 (Correctness Reward, 权重 90%)： 判断 \boxed{} 内的答案是否正确。若正确，则奖励为 1，否则为 0。这是模型核心能力的直接体现。

2.2 核心算法：改进的 GRPO

我选择 GRPO (Group Relative Policy Optimization) 作为核心强化学习算法。值得注意的是，我进行了一项关键改进：移除了传统的 KL 散度正则项。

这一决策的理论依据参考了 GLM4.5 技术报告及字节跳动的 DAPO 算法。移除 KL 约束可以避免对模型输出空间的过度限制，赋予模型更大的探索自由度，尤其是在需要创造性解题步骤的数学领域，这种探索能力至关重要。

2.3 关键训练参数

我在 Big-Math-Processed 数据集上进行训练，并在 GSM8K 测试集上进行验证。训练时的全局 batch size 设为 264，其中 mini-batch 为 72，每块 GPU 的 micro-batch 为 4。输入序列的最大长度为 2048 tokens，输出序列的最大长度为 512 tokens；

在算法设计上，我采用 Group Relative Policy Optimization (GRPO) 作为优势函数估计方法；基础模型选用 OpenSeek-Small-v1-SFT，优化器学习率设为 1e-5。训练过程中采用 BF16 精度以降低显存消耗；在并行与推理策略方面，使用 vLLM 推理引擎进行 rollout，每个 prompt 生成 6 个样本。

训练环境为 单机 6 卡 GPU。总训练轮数为 15，在训练开始前和此后每 10 步进行一次验证；模型 checkpoint 每 200 步保存一次，训练日志通过 TensorBoard 记录。

3. 实验结果与分析

我通过 TensorBoard 对训练过程中的关键指标进行了全方位监控。截至撰写报告，已训练 8000+steps，线上分数 13.02。各项数据显示，本次基于 GRPO 的强化学习训练过程稳定、收敛趋势明确，并成功提升了模型的数学推理能力。

3.1. 核心性能指标：验证集 Reward 稳步提升

模型在验证集上的 Pass@1 的 Reward 是衡量其泛化能力的核心指标。如图 1 所示，该指标呈现出清晰且持续的上升趋势：

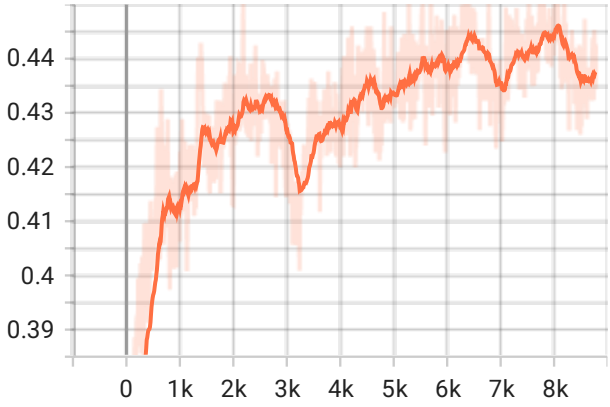

图 1 验证集 Reward

趋势分析： 在约 8,000 个训练步骤中，模型的 Pass@1 准确率从初始的约 7% 稳步攀升至峰值的 44.7%。这一显著的性能增益直接证明了我采用的奖励函数和 GRPO 算法对提升模型数学能力是行之有效的。

收敛状态： 在训练后期（6,000 步之后），曲线斜率逐渐放缓，表明模型性能趋于收敛，但仍有微小的提升空间。

3.2. 奖励函数与模型行为分析

奖励信号是驱动模型优化的直接动力。对奖励相关指标的分析可以揭示模型的学习动态。

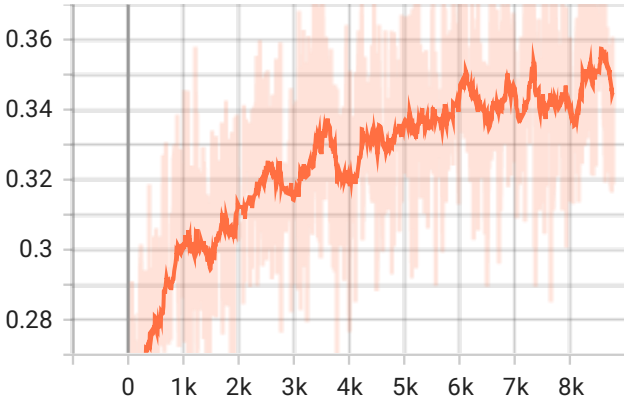

图 2 Critic 接收到的平均奖励 (rewards/mean)

奖励与准确率的强相关性： 如图 2 所示，模型获得的平均奖励从 0.28 持续增长到 0.354。这一趋势与验证集准确率的增长高度同步，表明模型正朝着我设定的奖励目标（即生成格式正确且答案准确的解法）进行有效优化。

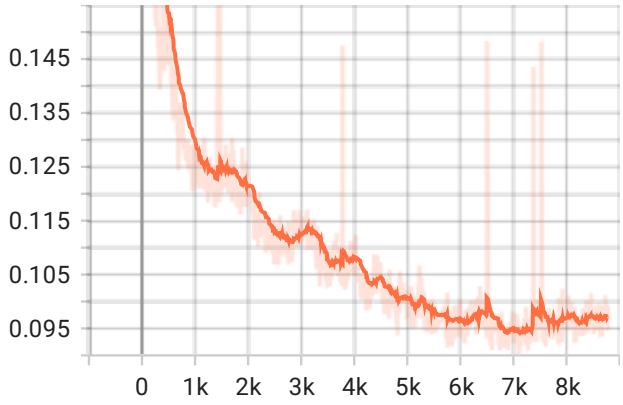

图 3 Actor 策略熵的变化

模型策略的收敛性： Entropy 指标衡量了模型输出策略的随机性。图 3 显示，熵值从约 0.18 平滑下降至 0.10 以下，这表明随着训练的进行，模型对其认为“最优”的解题路径变得越来越“自信”，策略正在从探索阶段过渡到利用阶段，是模型成功收敛的典型特征。

3.3 模型输出特征演化

我还观察到模型生成内容的特征在训练过程中发生了有趣的变化。

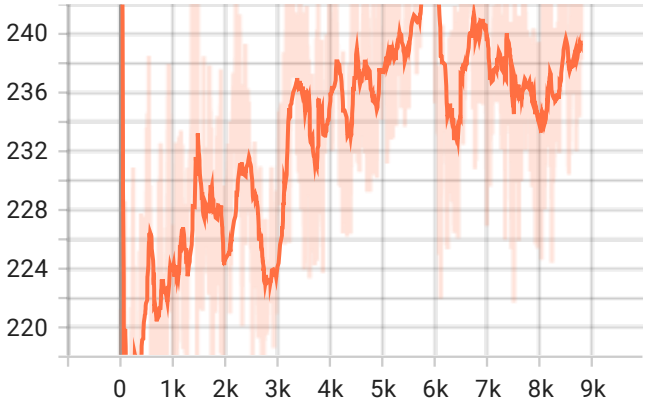

图 4 模型平均响应长度的变化

响应长度的增加： 如图 4 所示，模型输出的平均 Token 数量从约 220 增加到了 240 左右。这在一定程度上表明，模型可能正在学习生成更详尽、更具逻辑层次的解题步骤，而不是仅仅输出一个最终答案。这种行为与我 Prompt 中 "reason step by step" 的引导是一致的。

综上所述， 实验数据从外部性能（验证集准确率）、内部动机（奖励信号）和行为特征（策略熵、输出长度） 三个层面交叉验证了本次强化学习实验的成功。模型不仅在核心任务指标上取得了显著提升，其内部策略和外部行为也表现出符合预期的、积极的演化趋势。

4. 反思与展望

4.1. 当前工作的局限性

数据污染的初始疏忽： 训练初期的确未对数据污染问题给予足够重视，这提醒我数据的前期处理和验证是保证模型评估客观性的生命线。

超参调整不充分： 受限于时间和计算资源，rollout.n（每次探索的样本数）、max_new_tokens（生成长度）等关键超参未进行系统性寻优，当前配置可能并非最优解。

4.2. 未来工作展望

1. 算法层面：探索更前沿的 RLHF 算法

   尝试 DAPO/GSPO： GRPO 算法虽然有效，但业界已有更先进的替代方案。DAPO (Decoupled Clip and Dynamic sAmpling Policy Optimization) ， GSPO (Group Sequence Policy Optimization) 等新算法在训练稳定性、样本效率和最终性能上可能更具优势，值得尝试。

   引入 Process Reward Model (过程奖励模型)： 当前的奖励模型只关注最终结果。未来可以训练一个“过程奖励模型”，对解题的每一步进行打分。这能提供更密集的奖励信号，引导模型学习正确的推理路径，而不仅仅是蒙对答案。

2. 训练策略层面：引入课程学习 (Curriculum Learning)

从易到难的强化学习： 借鉴 GLM4.5 报告的启发，我可以设计一个从易到难的课程学习策略。首先，使用难度为 1-3 的简单问题进行第一轮强化，让模型快速建立信心和基础能力；然后，再使用难度为 4-5 的复杂问题进行第二轮强化，挑战模型的能力上限。

3.数据层面：精细化处理与增强

语义去重： 除了精确匹配，未来应采用基于 n-gram 或 Sentence-BERT 的语义相似度匹配方法，进一步筛查潜在的数据污染，确保评估的纯净性。

多样性增强： 针对当前数据集可能存在的解题思路单一问题，可以利用强大的外部大模型 API（如 Claude 4）进行数据增强。

4.模型与评估层面：

集成多个奖励模型： 单一奖励模型可能存在偏差。可以训练多个奖励模型，并在训练中对它们的输出进行集成（如投票或加权平均），以获得更鲁棒的奖励信号。

构建更全面的评估体系： 除了 Pass@1，还应引入 Pass@k、解题步骤正确率、不同知识点/难度下的细分准确率等指标，对模型能力进行更全面的画像。

后记

作为一名学生，一个单人团队，从初赛一路来到决赛，我很幸运。我很感谢赛事方提供的比赛以及算力支持，让我能去真正了解 LLM 的 CPT，RLHF，能去把理论去应用到实践中去；和各个业内大佬去竞争，去比赛，去交流，对我而言也是很大的提升。真的非常感谢能给我这样的机会！