

PR #2085 完整报告

THUDM/slime

fix(opd): score teacher logprobs at rollout temperature, not 0

合并时间: 2026-08-21 10:53

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2085>

执行摘要

- 一句话: 修复 OPD 教师温度硬编码为 0 的问题
- 推荐动作: 该 PR 不建议精读, 但其讨论过程值得关注, 特别是关于 SGLang 温度对 input logprob 影响的澄清, 以及最终引导至 PR#2325 的正确修复。

功能与动机

PR body 指出 OPD 的 KL 惩罚是学生 (rollout 采样) 与教师分布之间的差异, 而教师打分时硬编码 temperature=0 会导致贪心分布, 当 --rollout-temperature 不为 0 时, 惩罚与 rollout 策略不匹配。

实现拆解

1. 修改 slime/rollout/on_policy_distillation.py 中 reward_func 的 payload, 将 sampling_params 中的 temperature 从 0 改为 args.rollout_temperature。
2. 保持 max_new_tokens=0, 确保只对已有序列打分。
3. 该改动仅一行, 未涉及测试或其他模块。

关键文件:

- slime/rollout/on_policy_distillation.py (模块 蒸馏模块; 类别 source; 类型 core-logic; 符号 reward_func): 修改了 OPD 教师打分的 temperature 参数, 使之与 rollout 温度对齐, 但最终被发现为 no-op。

关键符号: reward_func

关键源码片段

[slime/rollout/on_policy_distillation.py](#)

修改了 OPD 教师打分的 temperature 参数, 使之与 rollout 温度对齐, 但最终被发现为 no-op。

```
async def reward_func(args, sample, **kwargs):
    # 构造评分请求 payload
    # 原实现 temperature 固定为 0, 意图是贪心分布;
    # 但 OPD 的 KL 惩罚需要教师分布与学生采样分布对齐,
    # 因此尝试改为 args.rollout_temperature。
```

```
# 注意: SGLang 的 temperature 只影响 output logprob,
# 对 input logprob (此处使用) 无效, 故实际为 no-op。
payload = {
    "input_ids": sample.tokens,
    "sampling_params": {
        "temperature": args.rollout_temperature, # 原为 0
        "max_new_tokens": 0, # 仅对已有序列打分
        "skip_special_tokens": False,
    },
    "return_logprob": True,
    "logprob_start_len": 0,
}
# ... 发送请求并返回 logprob
```

评论区精华

kkyyxhll 指出该修改实际上是 no-op, 因为 SGLang 的 temperature 只影响 output logprobs, 而对 input logprobs 无影响 (OPD 使用 max_new_tokens=0, 只取 input logprobs)。EazyReal 承认这一点, 但坚持认为应匹配教师分布 p_T 而非 p_1 , 并提出了更大的 SGLang 输入 logprob 修改方案 (PR#2325)。

- OPD 教师温度修改是否为 no-op (correctness): 确认该改动无效, 需通过修改 SGLang 输入 logprob 的温度支持来实现真正对齐。
- 正确修复方案 (design): 接受 PR#2325 作为正确修复, 本 PR 关闭。

风险与影响

- 风险: 该改动本身为 no-op, 无技术风险。但若直接合并, 可能造成误导, 让人误以为温度已对齐, 而实际未生效。正确修复需要修改 SGLang 的 input logprob 计算逻辑, 风险较高, 需回归验证。
- 影响: 对用户和系统无实际影响, 因为改动未生效。但对团队而言, 明确了 OPD 教师温度对齐的正确路径, 需依赖于 SGLang 的 input logprob 温度支持。
- 风险标记: 无效修改, 需求未实现

关联脉络

- PR #2319 fix(opd): score teacher logprobs at rollout temperature, not 0: kkyyxhll 在评论中提出新 PR, 认为本 PR 无效, 建议正确实现。
- PR #2325 OPD 教师温度正确修复 (草稿): EazyReal 提出通过修改 SGLang input logprob 温度支持来实现真正对齐, 并带有回归测试。