

# PR #1298 完整报告

radixark/miles

[OPD] Per-position teacher scoring (sparse top-k) + kaixih's robustness fixes

合并时间: 2026-08-03 15:27

原文链接: <http://prhub.com.cn/radixark/miles/pull/1298>

## 执行摘要

- 一句话: OPD top-k 打分稀疏化, 响应降 22.9x
- 推荐动作: 值得精读, 尤其 `_per_position_ids` 的布局对齐设计 (prompt 空槽 + 绝对位置索引复用 `_trim_input_field`)、`parse_teacher_urls` 的 fail-fast 风格、以及 reward 日志摘要模式。建议跟进 review 遗留问题: 修复 `_post_json` 的 None 超时语义、确认 `--sglang-router-request-timeout-secs` 注册情况; 若计划上线 `--opd-topk-per-position`, 先确认 teacher/student 均为 patched 版本, 并在小规模 rollout 上比对 flat 与 per-position 的 reward 一致性。

## 功能与动机

OPD top-k teacher 评分存在二次方数据体积问题: PR body 明确指出当前实现把 " 全局 union 广播到每个位置 ", 导致响应为  $O(\text{response\_len} \cdot |\text{union}|)$  JSON, "dominated by client-side parse + map-building (the rollout-tail CPU bottleneck seen on B200)". 目标是借助 `sgl-project/sglang#27421` 的 per-position 字段把响应降到  $O(\text{response\_len} \cdot k)$ 。同时吸收 kaixih 的 `opd-qwen3-b200-fixes` 中正交的稳定性修复: 打分请求不能无限挂起、日志不能整 dump 嵌套 logprob dict; 第三个 commit 还引入多 teacher 路由, 使同一训练批次可按任务路由到专精 teacher。

## 实现拆解

实现按 5 步推进:

1. 打通稀疏打分协议: `_score_payload` 扩展 `token_ids_positions` 参数, 优先写入 `token_ids_logprob_positions`, 与旧的扁平 `token_ids_logprob` 互斥; 新增 `_per_position_ids` 把按 response 位置组织的 top-k 条目转成按绝对输入位置索引的 id 列表 (prompt 段用空列表占位), 保证与 `logprob_start_len=0` 及既有 `_trim_input_field` 的 `values[1:][-response_length:]` 取位方式一致。这是 per-position 正确性的核心: response 位置 `r` 的 id 列表必须落在绝对索引 `prompt_len + r`。
2. 重写 `reward_func` 打分流程: 先经 `_teacher_url_for_sample` 解析本样本的 teacher 端点 (未配置路由时回退 `--rm-url`); 再依 `--opd-topk-per-position` 在 per-position 与 flat 双路径间选择; teacher 与 student 两侧 (`STUDENT_ON_TEACHER_STRATEGIES`) 的打分都支持新布局; 所有 `_post_json` 调用统一带上 per-request `timeout_secs`。

3. 多 teacher 路由: `parse_teacher_urls` 只按第一个 = 切分 `NAME=URL`, 保留 URL 内 query 参数 (如 `?tag=a=b`), 对格式错误、重复名字直接抛错;  
`_teacher_url_for_sample` 按 `sample.metadata[--opd-teacher-key]` 路由, 缺失或未知名字回退保留名 `default`, 无 `default` 则报错, 奉行 " 静默蒸馏错 teacher 比直接失败更糟 " 的原则。
4. 参数与启动校验: `miles/utils/arguments.py` 注册 `--opd-topk-per-position`、`--opd-teacher-urls`、`--opd-teacher-key`; `miles_validate_args` 增加组合校验 (`--opd-teacher-urls` 仅限 `--opd-type=sglang`; 未开 `--use-opd` 时禁止设置), 并在启动期 fail-fast 解析路由表 (局部 import 避免 `miles.utils` 反向依赖 rollout)。
5. 日志与配套: `miles/rollout/sglang_rollout.py` 新增 `_len_or_value`、`_sample_text_preview`、`_reward_log_summary`, first/finish rollout 日志改为参数化格式字符串 + 形状摘要; `tests/fast/rollout/test_on_policy_distillation.py` 新增 12 个用例覆盖 padding 布局、双路径 payload、URL 解析与路由回退 (注册为 stage-a-cpu CI); 新增 `run-qwen3-8B-opd-multi-teacher.sh` 示例 (math teacher Qwen3-32B + code teacher Qwen3-Coder-30B-A3B) 及 `docs/advanced/on-policy-distillation.md`、示例 README 更新。

关键文件:

- `miles/rollout/on_policy_distillation.py` (模块 OPD 打分; 类别 source; 类型 core-logic; 符号 `parse_teacher_urls`, `_teacher_url_for_sample`, `_score_payload`, `_per_position_ids`): PR 核心: per-position 稀疏打分 (`_per_position_ids`、`_score_payload`)、多 teacher 路由 (`parse_teacher_urls`、`_teacher_url_for_sample`)、请求超时 (`_post_json`)、`reward_func` 双路径重写, 全部集中在此。
- `miles/rollout/sglang_rollout.py` (模块 rollout 主控; 类别 source; 类型 observability; 符号 `_len_or_value`, `_sample_text_preview`, `_reward_log_summary`): rollout 主循环日志改造: 用 `_reward_log_summary` 等摘要替代整 dump 嵌套 logprob, 避免 OPD 打分 reward 撑爆日志。
- `miles/utils/arguments.py` (模块 参数解析; 类别 source; 类型 configuration; 符号 `add_on_policy_distillation_arguments`, `miles_validate_args`): 新增 `--opd-topk-per-position`、`--opd-teacher-urls`、`--opd-teacher-key` 参数及 `miles_validate_args` 组合校验 (fail-fast 解析路由表)。
- `tests/fast/rollout/test_on_policy_distillation.py` (模块 OPD 测试; 类别 test; 类型 test-coverage; 符号 `test_per_position_ids_pads_prompt_and_keeps_response_order`, `test_score_payload_routes_per_position_vs_flat`, `test_parse_teacher_urls_pares_names_and_keeps_equals_in_url`, `test_parse_teacher_urls_rejects_malformed_entries`): 新增 12 个用例覆盖 per-position padding 布局、双路径 payload 路由、teacher-urls 解析与回退语义, 是验证布局对齐的关键测试。
- `examples/on_policy_distillation/run-qwen3-8B-opd-multi-teacher.sh` (模块 示例脚本; 类别 other; 类型 entrypoint): 新增多 teacher OPD 端到端示例: Qwen3-32B math teacher + Qwen3-Coder-30B code teacher, 按 `metadata opd_teacher` 标签路由, 验证新功能真实可用。

- docs/advanced/on-policy-distillation.md (模块 OPD 文档; 类别 docs; 类型 documentation) : 补充 --opd-teacher-urls、--opd-teacher-key 参数表与 Multi-Teacher Routing 章节, 说明共享 tokenizer 约束和 default 回退语义。
- examples/on\_policy\_distillation/README.md (模块 示例文档; 类别 docs; 类型 documentation) : 示例目录索引新增 multi-teacher 脚本说明。

关键符号: parse\_teacher\_urls, \_teacher\_url\_for\_sample, \_score\_payload, \_per\_position\_ids, \_post\_json, reward\_func, \_len\_or\_value, \_sample\_text\_preview, \_reward\_log\_summary

## 关键源码片段

### miles/rollout/sglang\_rollout.py

rollout 主循环日志改造: 用 `_reward_log_summary` 等摘要替代整 dump 嵌套 logprob, 避免 OPD 打分 reward 撑爆日志。

```
def _reward_log_summary(reward: Any) -> Any:
    """把 OPD 打分返回的嵌套 logprob dict 摘要成形状/长度, 替代整 dump。
```

```
    非 dict 奖励走 _len_or_value; dict 则逐层收敛: 外层键 -> 子 dict 的
    keys 列表 + meta_info 关键字段的长度摘要, 控制日志体积。
```

```
    """
```

```
    if not isinstance(reward, dict):
        return _len_or_value(reward)
```

```
    summary: dict[str, Any] = {}
    for key, value in reward.items():
        if not isinstance(value, dict):
            summary[key] = _len_or_value(value)
            continue
```

```
    entry: dict[str, Any] = {"keys": list(value.keys())}
    meta_info = value.get("meta_info")
    if isinstance(meta_info, dict):
        # 只保留高频字段的形状信息。
        entry["meta_info"] = {
            meta_key: _len_or_value(meta_info[meta_key])
            for meta_key in (
                "id",
                "finish_reason",
                "prompt_tokens",
                "weight_version",
                "input_token_logprobs",
                "input_token_ids_logprobs",
                "input_top_logprobs",
            )
            if meta_key in meta_info
        }
```

```
summary[key] = entry
return summary
```

```
def _len_or_value(value: Any) -> Any:
    # 容器型字段用 {type, len} 摘要，标量原样返回。
    if isinstance(value, (dict, list, tuple, str)):
        return {"type": type(value).__name__, "len": len(value)}
    return value
```

## 评论区精华

review 的核心交锋集中在 `_post_json` 的超时语义。gemini-code-assist[bot] 以高优先级指出：`timeout_secs` 为 `None` 时 `aiohttp.ClientTimeout(total=None)` 会禁用 aiohttp 默认的 5 分钟超时，请求可能无限挂起，并给出 " 仅在非 `None` 时传 `timeout`" 的 suggestion；合入者 Shi-Dong 回复 'Please address this.' 要求处理，但 head 代码中该建议未被采纳，属已批准却遗留的高风险点。第二个高优先级意见是 `--sglang-router-request-timeout-secs` 未在 `miles/utils/arguments.py` 注册（commit 消息称其为 existing 参数，head 快照中未见注册代码），若属实该超时保护实际不可用。另有 medium 级 `prompt_len` 防御式计算建议，以及 Shi-Dong 的命名 nit（测试导入的 `_compute_topk_reverse_kl`、`_per_position_ids` 等建议去掉前导下划线），两者均未处理。

- `_post_json` 传 `None` 超时会禁用 aiohttp 默认超时 (correctness): head 代码未采纳建议，`_post_json` 仍保留原写法；PR 已合并，该风险留存。
- `--sglang-router-request-timeout-secs` 参数是否注册存疑 (correctness): head 快照的 `arguments.py` 中未见注册代码，需人工确认；若属实，超时保护实际不可用。
- `prompt_len` 计算的防御式处理 (correctness): 未采纳；现有调用上下文由 `Sample` 类型保证字段非空，风险较低。
- 测试导入的私有函数命名 (style): 未处理，维持 `_` 前缀；属于风格建议，不影响功能。

## 风险与影响

- 风险：
  1. 超时语义回退（高）：`_post_json` 在 `timeout_secs=None` 时构造 `ClientTimeout(total=None)`，会禁用 aiohttp 默认超时，使 " 防挂起 " 目标可能落空甚至引入无限等待；Gemini 已指出、Shi-Dong 要求处理但合并时 head 代码未改。
  2. 外部 `sglang patch` 依赖：`--opd-topk-per-position` 依赖 `sgl-project/sglang#27421` 的 patch；若 teacher 与 student 服务器 patch 版本不一致，未知字段可能被旧版忽略或报错。默认关闭缓解了未部署场景，但开启后没有运行时检测。
  3. 布局对齐耦合：`_per_position_ids` 依赖 `sglang` 侧按绝对位置、chunked prefill 下对齐（`LogitsMetadata.extend_logprob_start_pos_cpu`）等实现细节；`miles` 侧单测只验证 padding 布局，端到端对齐由外部 PR 保证，若 `sglang patch` 有偏差，训练信号会静默错位（错误 token 的 logprob），且不易察觉。

4. 参数注册存疑: `--sglang-router-request-timeout-secs` 在 head 快照的 `arguments.py` 中未见注册代码, 若属实则命令行传参会报 `unrecognized arguments`, 超时保护实际不生效。
5. 日志格式变更: `first/finish rollout` 日志从完整 `reward dump` 变为摘要, 依赖日志全文的运维 / 排查脚本需适配; 多 `teacher` 路由启用后 `--rm-url` 被忽略, 用户需注意配置语义。
  - 影响: 训练效率: 开启新开关后, B200 上 OPD 打分客户端的解析 / 建 map CPU 开销显著降低 (响应实测 22.9x); 功能面: 支持任务级多 `teacher` 路由, 同一训练任务可混合 `math/code` 等专精 `teacher`, OPD 配方更灵活; 稳定性: 打分请求不再可能无限阻塞 rollout (前提是超时参数真正可用并修复 `None` 语义); 团队: 需要部署 patch 后的 `sglang` (`sglang-miles` 分支) 才能开启 `per-position`, 文档与示例已同步, 成本可控。整体影响集中在 OPD 训练路径, 默认关闭保证对存量任务零侵入。
  - 风险标记: `None` 超时禁用默认限制, 依赖外部 `sglang` patch, 待核实参数注册, 布局对齐耦合外部实现

## 关联脉络

- PR #994 feat/opd-ci-docs (stack base) : PR body 明确声明 Stacked on #994 (base 分支 feat/opd-ci-docs), 本 PR 的 `per-position` 打分与鲁棒性修复均在该 OPD 功能线上叠加。
- PR #1314 Multi-teacher routing: per-sample teacher selection via `--opd-teacher-urls`: 第三个 commit 将 #1314 的实现并入本分支 (提交消息 [OPD] [3/N] Multi-teacher routing...), 与本 PR 的 `per-position` 打分同属 OPD 系列改进。
- PR #27421 [logprob] Per-position token\_ids\_logprob for sparse OPD top-k scoring: 跨仓库依赖 (sgl-project/sglang) : `per-position` 路径的前置条件, 必须在 `teacher/student` 服务器部署该 patch 才能开启 `--opd-topk-per-position`。