# PR #36979 完整报告

sgl-project/sglang

[Test] Move `gpqa` and `aime25` onto sgl-eval, drop unused eval paths

合并时间：2026-08-30 08:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36979

## 执行摘要

- 一句话：gpqa/aime25 评测统一到 sgl-eval，删除 1800+ 行死代码
- 推荐动作：值得精读 run_eval.py 的改动：_run_sgl_eval 中 reasoning_effort/seed/repeat 的透传方式，是评测框架迁移时最容易踩的坑，本 PR 给出了正确处理；三 sigma 基线重测方法论与 CLI 参数清理清单也可直接复用。整体属于评测基础设施清理，不涉及推理核心路径，适合作为「评测统一化」类 PR 的参考模板。

## 功能与动机

PR body 明确说明了这次收敛的动机：gpqa 除本分支外，其余调用方（eval_accuracy_kit）早已通过 sgl-eval 计分，simple_eval_gpqa.py 是最后一条未迁走的路径；aime25 的两个调用方也已走 kit 的 sgl-eval 驱动，run_eval 分支只剩 CLI 入口可达。math 零调用方且其 equality checker 依赖实时 gpt-4-turbo key，维护成本高；longbench_v2 评测路径与四个 CLI 参数同样零调用（kl_test_utils.py 与 benchmark/datasets/longbench_v2.py 只是把数据集当长文本来源，不经过 run_eval）。核心诉求是让评测体系收敛到单一 sgl-eval 计分器，消灭多套计分逻辑并存带来的阈值漂移与维护负担。

## 实现拆解

整个变更以 python/sglang/test/run_eval.py 为枢纽，分为六步完成：

1. 路由改造：run_eval.py 中 gpqa 与 aime25 两个分支从「构造本地 Eval 对象」改为 return _run_sgl_eval("gpqa", args) 与 return _run_sgl_eval("aime25", args)，统一由 sgl-eval 的 mcq prompt + eval_mcq 计分器打分；同时删除 math、mgsm、longbench_v2 三个死分支。
2. 参数透传：_run_sgl_eval 新增 reasoning_effort 透传逻辑——gpt-oss 按 effort 档位分别计分，若不透传，三个档位会全部塌缩到服务模型的默认档位；这是本次迁移中最容易被忽略的正确性细节。
3. 重复采样固定：python/sglang/test/gpt_oss_common.py 中固定 repeat=1，因为 sgl-eval 的 gpqa 默认 n_repeats=8，会成倍放大运行时间与统计口径。
4. 死代码删除：删除 9 个无调用方文件，合计约 1770 行，详见下表。
5. 基线重测：按 sgl-eval 计分器重测 gpt-oss 系列阈值，取均值 3 sigma 之下（198 题，binomial sigma ≈ 0.034）。

6. 测试与 CI 配套：sm120 与 4gpu mxfp4（含 cp）测试阈值更新；CI 中经多轮 /rerun-test 验证，最终合并。

| 删除文件 | 行数 | 说明 |
| --- | --- | --- |
| python/sglang/test/simple_eval_longbench_v2.py | 344 | LongBenchV2Eval 全套实现 |
| test/manual/eval/validate_longbench_v2.py | 337 | 格式兼容性与管道验证脚本 |
| test/manual/eval/validate_longbench_v2_standalone.py | 306 | 独立版验证脚本 |
| test/manual/eval/test_longbench_v2_eval.py | 238 | LongBench-v2 单测 |
| test/manual/eval/longbench_v2_evaluation.md | 217 | LongBench-v2 评测指南 |
| python/sglang/test/simple_eval_aime25.py | 124 | AIME25Eval 实现 |
| python/sglang/test/simple_eval_gpqa.py | 97 | GPQAEval 实现 |
| python/sglang/test/simple_eval_math.py | 77 | MathEval 实现 |
| test/manual/core/test_gpt_oss_1gpu.py | 31 | 从未注册进 CI 的手动测试 |

关键文件：

- python/sglang/test/run_eval.py（模块 评测入口；类别 test；类型 core-logic；符号 _run_sgl_eval, run_eval）：评测入口核心改动：gpqa/aime25 分支改为 _run_sgl_eval 调用，删除 math/mgsm/longbench_v2 分支与 4 个 LongBench 专属 CLI 参数，并在 _run_sgl_eval 中新增 reasoning_effort 透传，是本次迁移与清理的枢纽。
- python/sglang/test/simple_eval_gpqa.py（模块 评测脚本；类别 test；类型 deletion；符号 GPQAEval）：被删除的自研 GPQA 评测实现（97 行），其功能已由 sgl-eval 的 mcq prompt + eval_mcq 计分器接管，删除后避免两套计分口径并存。
- python/sglang/test/simple_eval_aime25.py（模块 评测脚本；类别 test；类型 deletion；符号 AIME25Eval, normalize_aime_answer）：被删除的自研 AIME25 评测实现（124 行），含 QUERY_TEMPLATE 与 normalize_aime_answer 归一化逻辑；两个既有调用方已走 sgl-eval 驱动。
- python/sglang/test/simple_eval_longbench_v2.py（模块 评测脚本；类别 test；类型 deletion；符号 LongBenchV2Eval, format_longbench_v2_question, extract_longbench_v2_answer, _load_dataset）：被删除的最大评测实现（344 行），包含官方模板格式化、答案抽取、HF/ 本地多来源加载与长度过滤；对应 4 个 CLI 参数一并移除。
- python/sglang/test/simple_eval_math.py（模块 评测脚本；类别 test；类型 deletion；符号 MathEval）：被删除的 MATH 评测实现（77 行），零调用方且 equality checker 依赖实时 gpt-4-turbo key，维护成本高。

- python/sglang/test/gpt_oss_common.py（模块 测试基线；类别 test；类型 test-coverage；符号 BaseTestGptOss）：gpt-oss 系列测试公共基类：固定 repeat=1，规避 sgl-eval gpqa 默认 n_repeats=8 带来的重复采样与运行时间膨胀。
- test/registered/models_e2e/test_gpt_oss_sm120.py（模块 测试基线；类别 test；类型 test-coverage）：更新 sm120 三档基线阈值（0.48/0.39/0.26），并记录 high 档因 max_tokens=4096 截断 70% 答案的既有问题。
- test/registered/cp/test_gpt_oss_4gpu_mxfp4_cp.py（模块 测试基线；类别 test；类型 test-coverage）：更新 4-gpu mxfp4 cp 基线阈值（0.586 → 0.50），该测试在 CI 中经 4 次重跑才通过，是阈值贴近 3 sigma 抖动的直接证据。

关键符号：_run_sgl_eval, run_eval, GPQAEval, AIME25Eval, MathEval, LongBenchV2Eval, normalize_aime_answer

## 关键源码片段

### python/sglang/test/run_eval.py

评测入口核心改动：gpqa/aime25 分支改为 _run_sgl_eval 调用，删除 math/mgsm/longbench_v2 分支与 4 个 LongBench 专属 CLI 参数，并在 _run_sgl_eval 中新增 reasoning_effort 透传，是本次迁移与清理的枢纽。

```
def _run_sgl_eval(eval_name: str, args) -> dict:
    # 前半段负责组装 sgl-eval 命令（模型、端口、采样参数等），
    # 这里只摘录与本次变更相关的尾部参数透传逻辑。

    # seed 默认不设置；只有 temperature > 0 的采样类调用方才需要显式透传。
    if getattr(args, "seed", None) is not None:
        cmd += ["--seed", str(args.seed)]

    # gpt-oss 按 reasoning_effort 档位分别计分。若不透传该参数，
    # 三个档位会全部塌缩到服务模型的默认档位，导致高低档基线失真。
    if getattr(args, "reasoning_effort", None) is not None:
        cmd += ["--reasoning-effort", str(args.reasoning_effort)]

    if getattr(args, "repeat", None) is not None:
        cmd += ["--n-repeats", str(args.repeat)]

    # 限制生成长度，避免长思考模型拖慢评测。
    # ...


def run_eval(args):
    # ...
    elif args.eval_name == "gpqa":
        # gpqa 改由 sgl-eval 计分（NeMo-Skills 的 mcq prompt + eval_mcq 计分器），
        # 调用方的阈值必须针对新计分器重新测量，不能沿用旧口径。
        return _run_sgl_eval("gpqa", args)
    # ...
```

```
elif args.eval_name == "aime25":
    # aime25 的两个既有调用方已走 eval_accuracy_kit 的 sgl-eval 驱动，
    # 这里把 run_eval CLI 分支也统一过去，随后删除 simple_eval_aime25.py。
    return _run_sgl_eval("aime25", args)
```

## 评论区精华

该 PR 没有任何人工 review 评论（review_comments_count = 0），5 条 issue 评论均为 mintlify 预览 bot 与 /rerun-test 机器人记录，核心讨论实际沉淀在 PR body 的基线重测说明与 rerun 记录里：

> test_gpt_oss_4gpu_mxfp4_cp.py 需要 4 次重跑才通过；阈值压到均值 3 sigma 之下（198 题，binomial sigma ≈ 0.034）。
>
> sm120 的 high 档得分低于 low 档，是因为 max_tokens=4096 截断了 70% 的答案（medium 为 34%），不是模型退化——与既有 # TODO 4k is still not enough 注释描述的是同一现象。

这说明作者对「新计分器下旧阈值不可沿用」有清晰认知，但 3 sigma 之下仍出现多次重跑，阈值紧贴统计下限的抖动风险被接受。

- gpt-oss 基线阈值稳定性与 CI 重跑 (testing): 阈值重测方法合理（3 sigma 之下），但 cp 测试多次重跑说明统计下限仍有偶发失败风险，作者接受该风险并以重跑通过收尾。

## 风险与影响

- 风险：
  1. 计分器语义切换：gpqa/aime25 从本地正则抽答改为 sgl-eval 的 mcq prompt + eval_mcq 计分器，任何沿用旧阈值的外部调用方都会误判红 / 绿；本 PR 只重测了 gpt-oss 系列，覆盖范围外仍有风险。
  2. CLI 破坏性删除：run_eval.py 删除了 --dataset-path、--categories、--max-context-length、--min-context-length 四个参数，依赖这些参数的外部脚本会在运行时直接报错。
  3. 统计口径变化：gpt_oss_common 固定 repeat=1 后，单次采样方差变大；阈值又贴 3 sigma 之下，从 CI 中 cp 测试 4 次重跑可见偶发失败并非理论风险。
  4. 删除连锁反应：被删文件若存在未发现的 import（如外部仓库或未注册测试）会触发 ImportError；PR body 已声明 math/mgsm/longbench_v2 零调用方，且 kl_test_utils.py 与 benchmark/datasets/longbench_v2.py 走数据集直读路径不受影响。
  5. 既有问题保留：sm120 high 档 max_tokens=4096 截断 70% 答案的问题未在本次修复，读测试结果时需注意口径。 - 影响：对评测使用者：gpqa 与 aime25 的结果口径随计分器切换而变化，阈值需要按 sgl-eval 重测；run_eval CLI 的 LongBench 参数被移除。对代码库：净删约 1800 行，评测入口从多套 simple-eval 实现收敛为 sgl-eval 为主，遗留维护负担显著下降。对 CI：gpt-oss 系列测试阈值更新，减少因口径不匹配导致的误红。对团队：确立「新评测优先走 sgl-eval」的演进方向，run_eval.py 中保留的 mmlu 分支也注明了继续保留的原因（ascend eval 依赖其 subject2category 映射），说明这是一次有意识的收敛而非一刀切。 - 风险标记：评测计分器切换，CLI 参数破坏性删除，

阈值贴近 3 sigma，1800+ 行死代码删除

## 关联脉络

- 暂无明显关联 PR