

PR #2257 完整报告

radixark/miles

test(ci): disable router-equivalence and MiniMax-M2.7 E2E

合并时间: 2026-08-08 05:28

原文链接: <http://prhub.com.cn/radixark/miles/pull/2257>

执行摘要

- 一句话: 禁用两个过时 SGLang E2E, 释放 GPU CI 容量
- 推荐动作: 这是一个小而清晰的 CI 用例管理 PR, 值得花费几分钟了解 `register_cuda_ci` 的 `disabled` 语义: 通过保留注册 + 禁用原因字符串的方式管理退役用例, 比直接删除文件更利于审计与复启。若有类似“淘汰测试套件”的需求, 可作为团队范本参考。

功能与动机

PR body 明确说明: stage-c-4-gpu-h200 通道仍调度两个不再值得占用 GPU 时间的 SGLang 套件——router equivalence 测试在测试已弃用的 Miles Router, MiniMax-M2.7 测试针对已弃用的模型。作者选择把它们标记为 `disabled` 以保留注册可见性, 同时不消耗 CI 容量, 并将 `est_time` 修正为观测运行时长, 以便重新启用时调度准确。

实现拆解

变更入口为两个 E2E 测试文件顶部的 `register_cuda_ci` 注册调用。

1. 在 `tests/e2e/sglang/test_r3_router_equivalence.py` 中, 将 `register_cuda_ci` 由单行调用改为多行, 新增 `disabled="Miles Router is deprecated."`, 并把 `est_time` 从 1000 修正为观测到的 1100。
2. 在 `tests/e2e/sglang/test_session_server_multi_role/test_minimax_m27.py` 中同样新增 `disabled="MiniMax-M2.7 is deprecated."`, 把 `est_time` 从 600 修正为观测到的 800。
3. 复用现有 `run_suite.py` 的 `disabled` 机制 (`run_suite.py:79` 将 `disabled` 注册排除在 `enabled` 集合外, `run_suite.py:138` 以 `skipped` 报告原因), 无需改动调度器代码。
4. 两个文件均保留全套用例与 `register_ci_gate` 门禁注册, 仅停止调度; 现有 `test_run_suite.py` 断言 `disabled` 注册恒被分类为 `skipped`, 因此无需新增测试。

关键文件:

- `tests/e2e/sglang/test_r3_router_equivalence.py` (模块 路由等价; 类别 test; 类型 test-coverage): 禁用基于已弃用 Miles Router 的等价性测试注册, 并将 `est_time` 修正为观测值 1100, 是本 PR 的核心变更文件之一。
- `tests/e2e/sglang/test_session_server_multi_role/test_minimax_m27.py` (模块 会话测试; 类别 test; 类型 test-coverage): 禁用针对已弃用模型 MiniMax-M2.7 的多角色会话测试, 同时保留 TITO 指标门禁注册, 是本 PR 的另一核心变更文件。

关键符号: register_cuda_ci

关键源码片段

tests/e2e/sglang/test_r3_router_equivalence.py

禁用基于已弃用 Miles Router 的等价性测试注册，并将 est_time 修正为观测值 1100，是本 PR 的核心变更文件之一。

```
# 来自 tests/e2e/sglang/test_r3_router_equivalence.py
from tests.ci.ci_register import register_cuda_ci

# Miles Router 已弃用，等价性测试不再占用 GPU 时间。
# 通过 disabled 参数将其从 stage-c-4-gpu-h200 调度集合中移出
# (run_suite.py 会将 disabled 注册报告为 skipped)，
# est_time 修正为观测到的 1100 秒，便于日后复启时调度准确。
register_cuda_ci(
    est_time=1100,
    suite="stage-c-4-gpu-h200",
    labels=["sglang"],
    disabled="Miles Router is deprecated.",
)
```

tests/e2e/sglang/test_session_server_multi_role/test_minimax_m27.py

禁用针对已弃用模型 MiniMax-M2.7 的多角色会话测试，同时保留 TITO 指标门禁注册，是本 PR 的另一核心变更文件。

```
# 来自 tests/e2e/sglang/test_session_server_multi_role/test_minimax_m27.py
from tests.ci.ci_register import register_cuda_ci
from tests.ci.metric_history import register_ci_gate

# MiniMax-M2.7 模型已弃用，整个套件标记 disabled 后不再调度，
# 但注册仍然可见，方便追踪与复启；指标门禁 (TITO session mismatch rate)
# 一并保留，套件恢复后即可继续生效。
register_cuda_ci(
    est_time=800,
    suite="stage-c-4-gpu-h200",
    labels=["sglang"],
    disabled="MiniMax-M2.7 is deprecated.",
)
register_ci_gate(metric_key="rollout/tito_session_mismatch_rate/v1/assistant_text")
register_ci_gate(metric_key="rollout/tito_session_mismatch_rate/v2/assistant_text")
```

评论区精华

该 PR 没有任何 review 评论或讨论线程。PR body 的 Review Focus 提出“仔细核对两条 disabled 原因，确认意图是完全关闭而非迁移到 nightly 节奏”，属于作者自设的审阅检查点，最终由作者自行合入。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 覆盖收缩：Miles Router 与 MiniMax-M2.7 两条回归线永久停跑，若相关组件仍被引用，回归问题将不再被 CI 捕获。
 - TITO 门禁指标断源：test_minimax_m27.py 的 register_ci_gate 仍然注册 rollout/tito_session_mismatch_rate 指标，套件禁用后该指标的 metric history 将不再产生新数据，若 CI 侧对该 gate 有强制要求，需确认无空窗告警。
 - disabled 参数依赖：register_cuda_ci 须已支持 disabled 关键字，本 PR 依赖 run_suite.py 既有实现；若参数名不匹配会导致导入期报错（当前已合并，风险已消除）。
 - 影响：影响集中在 CI 容量与回归覆盖：释放 stage-c-4-gpu-h200 通道的两段 GPU 时间（合计约 1900 秒 / 次），降低调度成本；对用户无感知，对团队而言减少两个维护项，但相应失去两条 E2E 回归保护。影响范围局限于 tests/e2e/sglang 下的两个注册点。
 - 风险标记：测试覆盖收缩，TITO 门禁指标断源，CI 注册参数变更

关联脉络

- PR #2218 ci: version TITO metrics by session server: 本 PR 分支名 jiajun/tito-metrics-by-session-server 与其同名，且被禁用的 MiniMax-M2.7 套件正是注册 TITO metric gate 的会话测试，属于同一 TITO 指标与 session 测试演进线。
- PR #2230 fix(ci): disable MI300X runner jobs: 同样是禁用 CI 用例以释放 GPU 资源，展示了仓库周期性裁剪 CI 用例的维护模式。