# PR #2214 完整报告

radixark/miles

fix(ci): calibrate lora E2E estimates from nightly runs and halve GLM5 lora matrices

合并时间：2026-08-14 06:28

原文链接：http://prhub.com.cn/radixark/miles/pull/2214

## 执行摘要

- 一句话：校准 lora E2E 时长估计，GLM5 矩阵减半
- 推荐动作：值得精读，尤其适合 CI 维护者和 *e2e* 测试作者。亮点在于：1）基于实测数据校准估计值的完整流程（公式、桶取整、跨夜验证）；2）通过互补对角线在保持全覆盖的前提下减半矩阵，并论证代码路径不相交的严谨性。可借鉴到其他耗时测试的优化中。

## 功能与动机

PR body 明确指出 *est_time* 远高于实测：*qwen3_5* 测试 *est*=3600 而实测 957–966s，GLM5 4 组合运行 1825–1829s 超过 1800s 默认超时，仅因 *est*=2400 抬高有效超时才通过。需要基于实测校准，并按照 *.claude/skills/ci-e2e-time-tune* 的公式 max(*elapsed*)×1.25 向上取整到 100s 桶。

## 实现拆解

1. 校准 *est_time*：依据 nightly runs 30926155662 和 30754322985 的完整日志，对每个测试取通过的尝试的最大耗时 ×1.25，向上取整到 100 秒桶。涉及 4 个文件：*test_qwen3_5_35b_a3b_lora_ci.py* 从 3600 调整为 1300，*test_inkling_small_4layer_lora_ci.py* 从 1800 调整为 800，GLM5.1 从 2400 调整为 1300，GLM5.2 从 2400 调整为 1200。
2. 缩减 GLM5 矩阵：将 *test_glm5_1* 和 *test_glm5_2* 的 _CONFIGS 从 4 个组合改为互补对角线（GLM5.1 保留 *tilelang+shared-outer+virtual-experts* 和 *megatron+per-expert+no-virtual-experts*；GLM5.2 保留 *megatron+shared-outer+virtual-experts* 和 *tilelang+per-expert+no-virtual-experts*），两个文件合起来仍覆盖全部 4 个 {dsa backend}×{lora layout} 单元。
3. 配套调整：*test_qwen3_5_35b_a3b_lora_ci.py* 的 labels 从 ["model-scripts"] 扩展为 ["megatron", "model-scripts", "lora"]，使其与其他 LoRA 测试一致，可能影响 CI 分区打包。验证方式为 run_suite.py --list-only 检查所有分区、证据 CSV 机械重算，并确认 6 个 lora 测试文件在证据运行 *head* SHA 与本分支 *base* 之间字节一致。

关键文件：

- *tests/e2e/megatron/model_scripts/test_glm5_2_744b_a40b_5layer_lora_ci.py*（模块 LoRA 测试；类别 test；类型 *test-coverage*；符号 _CONFIGS）：GLM5.2 5-layer LoRA 测试：*est_time* 2400→1200，_CONFIGS 从 4 组合缩减为 2 组合（

megatron+shared-outer+virtual-experts、tilelang+per-expert+no-virtual-experts），是矩阵减半的核心文件之一。

- tests/e2e/megatron/model_scripts/test_glm5_1_744b_a40b_6layer_lora_ci.py（模块 LoRA 测试；类别 test；类型 test-coverage；符号 _CONFIGS）：GLM5.1 6-layer LoRA 测试：est_time 2400→1300，_CONFIGS 从 4 组合缩减为 2 组合（tilelang+shared-outer+virtual-experts、megatron+per-expert+no-virtual-experts），与 GLM5.2 形成互补对角线的另一端。
- tests/e2e/megatron/model_scripts/test_inkling_small_4layer_lora_ci.py（模块 LoRA 测试；类别 test；类型 test-coverage）：Inkling Small 4-layer LoRA 测试：est_time 从 1800 校准为 800（实测 568s），仅 1 行改动，但体现数据驱动的校准方法论。
- tests/e2e/megatron/test_qwen3_5_35b_a3b_lora_ci.py（模块 LoRA 测试；类别 test；类型 test-coverage）：Qwen3.5-35B-A3B LoRA 测试：est_time 从 3600 校准为 1300（实测 957–966s），同时 labels 从 ["model-scripts"] 扩展为 ["megatron", "model-scripts", "lora"]，使其归入 LoRA 分组，影响 CI 分区选择。

关键符号：register_cuda_ci

## 关键源码片段

### tests/e2e/megatron/model_scripts/test_glm5_2_744b_a40b_5layer_lora_ci.py

GLM5.2 5-layer LoRA 测试：est_time 2400→1200，_CONFIGS 从 4 组合缩减为 2 组合（megatron+shared-outer+virtual-experts、tilelang+per-expert+no-virtual-experts），是矩阵减半的核心文件之一。

```
import os

from scripts.run_glm5_2_744b_a40b_lora import ScriptArgs, _prepare_download, _train
from tests.ci.ci_register import register_cuda_ci

import miles.utils.external_utils.command_utils as U

# Smoke test for scripts/run_glm5_2_744b_a40b_lora.py on the 5-layer toy (DSA cross-layer
# path, full rollout -> train -> save loop). Runs one diagonal of the MoE-expert LoRA
# matrix — {megatron + shared-outer + virtual-experts, tilelang + per-expert +
# no-virtual-experts} — and every combination must pass. The GLM-5.1 6-layer sibling runs
# the complementary diagonal, so nightly still covers all four {dsa backend} x {lora layout}
# cells across the pair. Functionality, not accuracy; 4 GPUs (TP=EP=4).

# 校准后的估计时长：实测约 896s，按 max(elapsed) x 1.25 向上取整到 100s 桶
register_cuda_ci(est_time=1200, suite="stage-c-4-gpu-h200", labels=["megatron", "model-scripts", "lora"])

# (name, dsa_attention_backend, experts_shared_outer_loras, virtual_experts_serving)
# 互补对角线：与 GLM-5.1 文件形成配对，避免重复运行相同组合
_CONFIGS = [
```

```
    ("megatron + shared-outer + virtual-experts", "megatron", True, True),
    ("tilelang + per-expert + no-virtual-experts", "tilelang", False, False),
]
```

### tests/e2e/megatron/model_scripts/test_glm5_1_744b_a40b_6layer_lora_ci.py

GLM5.1 6-layer LoRA 测试：est_time 2400→1300，_CONFIGS 从 4 组合缩减为 2 组合（tilelang+shared-outer+virtual-experts、megatron+per-expert+no-virtual-experts），与 GLM5.2 形成互补对角线的另一端。

```
import os

from scripts.run_glm5_1_744b_a40b_lora import ScriptArgs, _prepare_download, _train
from tests.ci.ci_register import register_cuda_ci

import miles.utils.external_utils.command_utils as U

# Smoke test for scripts/run_glm5_1_744b_a40b_lora.py on the 6-layer toy (full rollout ->
# train -> save loop). Runs one diagonal of the MoE-expert LoRA matrix — {tilelang +
# shared-outer + virtual-experts, megatron + per-expert + no-virtual-experts} — and every
# combination must pass. The GLM-5.2 5-layer sibling runs the complementary diagonal, so
# nightly still covers all four {dsa backend} x {lora layout} cells across the pair.
# Functionality, not accuracy; 8 GPUs (TP=EP=8).

# 校准后估计时长：实测约 985s，按公式向上取整到 100s 桶
register_cuda_ci(est_time=1300, suite="stage-c-8-gpu-h200", labels=["megatron", "model-scripts", "lora"])

# (name, dsa_attention_backend, experts_shared_outer_loras, virtual_experts_serving)
# 互补对角线：与 GLM-5.2 文件形成配对，覆盖另一半组合
_CONFIGS = [
    ("tilelang + shared-outer + virtual-experts", "tilelang", True, True),
    ("megatron + per-expert + no-virtual-experts", "megatron", False, False),
]
```

## 评论区精华

本 PR 无 review 讨论（0 条评论、0 条 review 评论），维护者 guapisolo 直接批准。PR body 中已详细论证矩阵缩减无覆盖损失，因此未引发讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：主要风险在于 测试覆盖缩减：GLM5 矩阵从 4 组合减到 2 组合，依赖'两维度互不相交'的论证（DSA backend 仅 train 侧，expert LoRA 布局不触碰 attention/qkv-format）。若未来代码路径发生变化使该假设失效，可能漏测。其次，est_time 下调后若 nightly 环境变慢（如资源争抢），可能触发超时；但 body 给出约 2 倍 headroom（1300 vs 896s）

，且两次 nightly 漂移 <2%，风险可控。另外 labels 变更可能影响 CI 分区选择，需观察后续打包行为。

- 影响：对用户无功能影响；对 CI 系统有直接收益：每晚节省约 1750 GPU 秒，lora 测试估计总时长从 12200s 降至 6600s，改善 LPT 分区打包效率。团队维护者受益于更准确的时长估计和更少的超时告警。影响范围仅限于 CI 配置与 *e2e* 测试定义，不改动任何生产代码。
- 风险标记：CI 时长估计下调，测试矩阵缩减，标签变更

## 关联脉络

- PR #2499 feat(ci): add weekly full-suite cadence: 同为 CI 基础设施演进，涉及 tests/ci 下的调度与分区策略，本 PR 的 est_time 校准直接影响分区打包效率。
- PR #2403 fix(ci): unblock ROCm fork PRs at checkout: 同为 CI 修复，关注 CI 流程稳定性，与本 PR 共同维护 CI 可靠性。