PR #22237 完整报告

sgl-project/sglang

[CI] Relax gpt-oss 4GPU accuracy threshold from 0.60 to 0.58

合并时间：2026-04-08 17:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22237

执行摘要

- 一句话：降低 GPT-OSS 4GPU 测试的准确度阈值，减少 CI 误报。
- 推荐动作：该 PR 变更简单直接，无需深入精读。值得关注的是其基于数据的阈值调整方法：通过分析历史运行数据（40 次调度运行）确定合理阈值，可作为类似测试稳定性优化的参考案例。

功能与动机

根据 PR body 中的分析，GPT-OSS 4GPU 测试在 B200 和 H100 runner 上均表现出 21% 的失败率（40 次调度运行），测试得分在 0.55-0.67 之间自然波动。0.60 的阈值过于严格，导致大量误报（例如链接中的失败示例）。降低阈值至 0.58 可在保留检测真实回归能力的同时，提供足够的容错空间。

实现拆解

仅修改一个测试文件中的两个测试函数：

1. test_bf16_120b：将 *expected_score_of_reasoning_effort* 的 "low" 值从 0.60 改为 0.58。
2. test_mxfp4_120b：进行相同的阈值调整。 所有变更集中在 test/registered/4-gpu-models/test_gpt_oss_4gpu.py 文件中，未涉及任何逻辑代码或配置变更。

关键文件：

- test/registered/4-gpu-models/test_gpt_oss_4gpu.py（模块 CI 测试）：唯一被修改的文件，包含两个测试用例的阈值调整，直接影响 CI 测试行为。

关键符号：test_bf16_120b, test_mxfp4_120b

评论区精华

该 PR 没有 review 评论或讨论，仅有一次提交历史显示阈值从 0.60 先降至 0.59，再进一步降至 0.58，表明作者基于数据持续优化阈值选择。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
  1. 无回归风险：仅修改测试阈值，不触及任何功能代码。
  2. 性能与安全：无影响。
  3. 兼容性：无影响。 唯一潜在风险是阈值降低可能掩盖真实性能回归，但 PR body 中展示的数据分析（得分趋势 0.55-0.67）支持 0.58 阈值仍能有效检测显著下降。
- 影响：影响范围有限：
  1. 对用户：无直接影响。
  2. 对系统：减少 CI 误报，提升测试稳定性，降低维护负担。
  3. 对团队：工程师将看到更可靠的 CI 结果，减少因测试波动导致的干扰。 影响程度为低，仅调整测试阈值。
- 风险标记：阈值调整可能掩盖真实回归

关联脉络

- PR #22346 [CI] Set RUNAI_STREAMER_MEMORY_LIMIT=0 for stage-b-test-1-gpu-small: 同属 CI 优化类别，关注测试稳定性和资源管理。
- PR #22292 [CI] Fix stage-b-test-1-gpu-large (0) timeout by reordering LoRA tests and using tokenizer from cache: 同属 CI 优化，通过调整测试顺序和加载策略解决稳定性问题。
- PR #22301 Only upload CUDA coredumps on test failure: 同属 CI 优化，减少资源浪费，提升效率。