PR #22297 完整报告

sgl-project/sglang

Revert "[CI] Update nightly test models for H200/B200 (#22288)"

合并时间：2026-04-08 08:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22297

执行摘要

- 一句话：撤销 H200/B200 夜间测试模型更新，恢复至先前配置。
- 推荐动作：该 PR 是一个简单的 revert 操作，建议快速审查以确保没有意外副作用，无需深入技术分析。但应关注是否后续有替代 PR 来解决原始问题，并监控 CI 稳定性。

功能与动机

PR body 未提供具体动机，但根据上下文推断，可能是由于 PR #22288 更新夜间测试模型后导致 CI 问题或需要临时回退。关联 Issue 为空，评论中只有机器人提示，无人类讨论表明具体原因。

实现拆解

本 PR 通过一个 commit 撤销了 commit e6652309（对应 PR #22288）。关键变更包括：删除 PR #22288 添加的 test/registered/8-gpu-models/test_glm_46.py 文件；修改其他四个测试文件（如 test_deepseek_v31.py、test_qwen35.py），移除 register_cuda_ci 调用和相关配置更改，使测试恢复为手动模式或旧有配置。例如，在 test_qwen35.py 中移除了 DP 相关参数，只保留 MTP 配置。

关键文件：

- test/registered/8-gpu-models/test_glm_46.py（模块 testing）：PR #22288 添加的 GLM-4.6 测试文件，在撤销中被删除，影响对新模型的支持。
- test/registered/8-gpu-models/test_qwen35.py（模块 testing）：恢复了 Qwen3.5 测试配置，移除了 DP 相关设置，减少测试变体，影响性能基准。
- test/registered/8-gpu-models/test_deepseek_v31.py（模块 testing）：移除了 register_cuda_ci 调用，恢复为手动测试模式，可能降低自动化测试效率。

关键符号：未识别

评论区精华

没有 review 评论或讨论，变更直接由作者合并，表明这是一个紧急或简单的回退操作。

- 暂无高价值评论线程

风险与影响

- 风险：撤销变更可能重新引入 PR #22288 试图解决的问题，具体风险包括：测试套件缺少对 GLM-4.6 模型的支持，可能导致新功能验证延迟；Qwen3.5 测试配置恢复为非 FP8 权重，可能影响性能基准的准确性；CI 测试效率可能降低，因为部分测试恢复为手动模式，减少自动化覆盖。
- 影响：直接影响 CI 测试基础设施，夜间测试将使用旧的模型集，对 H200/B200 平台的测试覆盖有负面影响。对终端用户无直接影响，但可能间接影响开发团队对新模型兼容性和性能的验证速度。
- 风险标记：测试覆盖减少，配置回退

关联脉络

- PR #22288 [CI] Update nightly test models for H200/B200: 本 PR 直接撤销了该 PR 的变更，恢复测试套件到先前状态。