PR #22288 完整报告

sgl-project/sglang

[CI] Update nightly test models for H200/B200

合并时间：2026-04-08 06:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22288

执行摘要

本 PR 更新了 H200/B200 GPU 的夜间测试套件，移除了 GLM-4.6、DeepSeek-V3.1 和 Qwen3-235B 的自动测试，将其转为手动执行，同时将 Qwen3.5 测试切换为 FP8 权重并添加 DP-attention 变体。变更旨在优化测试效率并反映最新模型支持，对内部 CI 流程有中等影响。

功能与动机

动机是精简夜间测试套件，移除已被新模型替代的旧测试（如 GLM-5 替代 GLM-4.6，DeepSeek-V3.2 替代 V3.1），并更新到更现代的模型版本。使用 FP8 权重可以提高测试的效率和代表性，适配 H200/B200 硬件。

实现拆解

实现涉及五个测试文件的修改：

- test_deepseek_v31.py、test_glm_46_fp8.py、test_qwen3_235b.py：移除 register_cuda_ci 调用，添加 "Manual-only" 注释，使其从 nightly 套件中取消注册。
- test_glm_46.py：被完全删除。
- test_qwen35.py：更新模型路径为 Qwen/Qwen3.5-397B-A17B-FP8，并引入新变体：python variants = [ModelLaunchSettings(..., variant="TP8+DP8"), ModelLaunchSettings(..., variant="TP8+DP8+MTP")]

评论区精华

本 PR 没有收到任何 review 评论，由作者直接合并，因此无讨论记录。

风险与影响

风险：

- 测试覆盖减少：移除自动测试可能降低回归检测能力。
- 权重变更风险：FP8 权重可能影响准确性测试基准。
- 新配置风险：DP-attention 变体可能未充分测试。

影响：

- 对用户无直接影响。

- 系统上，夜间测试套件运行更少测试，节省 CI 资源。
- 团队需调整测试策略，确保手动测试定期执行。

关联脉络

与历史 PR 关联：

- PR #22267：类似地将测试移至夜间套件，反映测试策略调整趋势。
- PR #21669：涉及 Qwen3.5 FP8 夜间性能测试，显示对该模型的持续关注。