PR #22140 完整报告

sgl-project/sglang

[Fix] Fix nightly tests

合并时间：2026-04-05 17:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22140

执行摘要

本 PR 修复了 SGLang 仓库夜间测试的多个不稳定和配置问题，包括取消不稳定测试注册、修复导入失败、调整精度阈值和更新测试模型，旨在提升 CI 流程的可靠性。

功能与动机

动机源于测试套件中的不稳定性和硬件兼容性问题。PR body 明确指出："Unregister test_lora_qwen3.py for cuda, since it's flaky and has been covered by some other tests... Fix register_custom_op import failure on hardware other than cuda... Lower threshold for nemotron 3 super nightly test... Skip qwen3 cp test on b200... Change dpsk-r1-fp4 nightly test to dpsk-v3-fp4 test"。

实现拆解

- deepseek_v2.py: 调整 import 语句，确保 register_custom_op 只在 CUDA 条件下导入，避免非 CUDA 硬件失败。python from sglang.srt.utils.custom_op import register_custom_op if _use_aiter:
- test_nvidia_nemotron_3_super_nightly.py: 将 GSM8K_BASELINE 阈值从 0.96 降至 0.935。python -GSM8K_BASELINE = 0.96 +GSM8K_BASELINE = 0.935
- test_qwen3_235b.py: 添加 @unittest.skipIf(is_blackwell_system(), ...) 装饰器跳过 B200 上的 cp 测试。
- test_lora_qwen3.py: 移除 register_cuda_ci 调用，取消 CUDA 测试注册。
- test_dpsk_v3_fp4_4gpu_perf.py: 重命名文件并更新配置，使用 DeepSeek V3 FP4 模型替换 R1 版本。

评论区精华

无 review 评论，PR 由作者直接合并，表明变更被视为小修复且无争议。

风险与影响

风险:

- 导入修复可能引入条件依赖错误，需验证非 CUDA 硬件路径。
- 阈值降低可能掩盖模型性能问题，需监控实际精度。

- 跳过 B200 测试可能遗漏硬件兼容性问题。
- 取消 LoRA 测试注册可能减少覆盖，需确认其他测试充分。

影响:

- 对用户无直接影响。
- 提升 CI 稳定性，减少资源浪费。
- 团队需适应测试配置变化。

关联脉络

与近期 PR 相关:

- PR #22137: 删除不稳定测试，类似策略减少 CI flaky。
- PR #22100: 调整测试阈值，同为修复不稳定测试。
- PR #22141: 优化 CI 工作流，提升测试效率。

这些 PR 共同反映了团队在持续改进测试可靠性和 CI 流程。