PR #21516 完整报告

sgl-project/sglang

[CI] Fix nemotron nvfp4 test estimated time

合并时间：2026-03-27 12:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21516

执行摘要

修复 NVIDIA Nemotron 3 Super NVFP4 模型 CI 测试的超时问题，通过将估计时间从 300 秒增加至 600 秒，确保测试稳定运行，避免不必要的 CI 失败。变更简单，影响范围仅限于该特定测试配置。

功能与动机

PR 动机源于 CI 测试超时失败，作者引用 GitHub Actions 运行链接（https://github.com/sgl-project/sglang/actions/runs/23623717495/job/68827833526?pr=20904）显示测试在 300 秒内未完成。因此，调整估计时间以避免超时，保证 CI 流程顺畅。

实现拆解

仅修改一个文件：test/registered/4-gpu-models/test_nvidia_nemotron_3_super_nvfp4.py。

关键代码变更：

```
- register_cuda_ci(est_time=300, suite="stage-c-test-4-gpu-b200")
+ register_cuda_ci(est_time=600, suite="stage-c-test-4-gpu-b200")
```

改动将 est_time 参数翻倍，为测试提供更多执行时间，不涉及其他逻辑调整。

评论区精华

无 review 评论或讨论，PR 由作者直接合并，表明变更无争议且被快速接受。

风险与影响

- 风险：增加估计时间可能掩盖测试执行缓慢的根本原因，如模型性能退化或资源竞争。但变更不涉及生产代码，直接风险低。
- 影响：仅影响特定模型 CI 测试，用户不受影响；CI 稳定性提升，但测试时间可能增加，需监控测试性能以防回归。

关联脉络

与近期 CI 相关 PR 如 #20942（更新依赖并启用测试）、#21492（修复 benchmark 空提示）和 #21501（扩展 CI 命令）类似，共同关注测试配置优化和 CI 可靠性。此外，动机链接提及 PR #20904，可能涉及相关测试上下文，显示团队对测试超时问题的持续处理。