PR #21485 完整报告

sgl-project/sglang

Remove redundant DeepSeek V3 FP4 PCG test

合并时间：2026-03-27 12:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21485

执行摘要

本 PR 删除了 DeepSeek V3 FP4 模型的一个冗余 PCG 测试类，并优化了 CI 估计时间，旨在减少资源浪费，提升测试效率，对系统无负面影响。

功能与动机

由于 PCG（Piecewise Cuda Graph）功能已默认启用，原有测试类 TestDeepseekV3FP4PiecewiseCudaGraph 与 TestDeepseekV3FP4 使用相同配置，造成重复执行，浪费 CI 时间和资源。PR body 中明确指出：'PCG is now enabled by default. TestDeepseekV3FP4PiecewiseCudaGraph runs the exact same config as TestDeepseekV3FP4 (both use default PCG) and is wasting CI time and resources.'

实现拆解

修改文件 test/registered/quant/test_deepseek_v3_fp4_4gpu.py：

- 删除冗余测试类：移除 TestDeepseekV3FP4PiecewiseCudaGraph 类及其所有方法（test_a_gsm8k 和 test_bs_1_speed），减少代码行数 67 行。
- 调整 CI 估计时间：将 register_cuda_ci 的 est_time 从 1500 秒改为 1200 秒，以更精准反映测试执行耗时。
- 优化代码风格：在保留的 TestDeepseekV3FP4.test_bs_1_speed 方法中，将未使用的变量 acc_length 替换为 _，提升代码可读性。

评论区精华

review 讨论仅聚焦于代码风格优化：

- gemini-code-assist[bot]建议：'The acc_length variable is assigned but not used. It's good practice to use _ for unused variables...'。
- 作者 mmangkad回复：'That's literally what the PR does'，确认变更已包含此改进。讨论无争议，已闭合。

风险与影响

- 风险分析：删除测试类可能误减覆盖度，但 PR 基于冗余事实，风险低；减少估计时间可能增加 CI 超时风险，但从 1500 秒降至 1200 秒是基于实际优化，风险可控。

- 影响分析：对用户无直接影响；系统 CI 执行更快，资源使用更高效；团队受益于更快的测试反馈，提升开发效率。

关联脉络

本 PR 与近期其他 CI 和测试优化 PR 关联，如：

- PR 21516：同样修复测试估计时间，优化 CI 资源使用。
- PR 21047：整合测试 mixins，减少冗余代码。 这表明仓库正在持续改进测试基础设施，以提升整体效率和可维护性。