PR #21779 完整报告

sgl-project/sglang

Reduce redundant speculative decoding CI tests

合并时间：2026-04-01 08:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21779

执行摘要

本次 PR 通过移除 EAGLE 推测解码测试中的冗余用例，将预估 CI 执行时间从 1661 秒减少至 850 秒，旨在加速 CI 流水线，变更仅影响测试文件，对核心功能无直接影响。

功能与动机

动机是减少 EAGLE 测试的 CI 执行时间，移除重复和跳过的测试用例。PR body 中明确说明："Reduce EAGLE test CI time by removing redundant/duplicate test cases." 具体目标包括删除已被覆盖的 TestEAGLERadixCache（已在 server-mode 测试中覆盖）和跳过的 TestEAGLEDraftExtend 类，并重构服务器测试以减少重复执行。

实现拆解

变更涉及两个测试文件：

- test_eagle_infer_a.py：删除 TestEAGLERadixCache 类（包含多配置测试）和 TestEAGLEDraftExtend 系列类，减少 265 行代码，将 register_cuda_ci 的 est_time 从 561 秒下调至 250 秒。
- test_eagle_infer_b.py：提取 TestEAGLEServerBasic 类为核心测试基类，仅保留 test_gsm8k 和 test_request_abort 方法；合并 TestEAGLEServerPageSizeTopkFA3 到 TestEAGLEServerExtend，est_time 从 1100 秒减少至 600 秒。

评论区精华

由于 review 评论为空，无实质性技术讨论。Issue 评论中作者多次使用 /rerun-test 命令验证 CI 通过（如 test_eagle_infer_b.py 和 test_eagle_infer_a.py），表明变更后测试执行正常，无争议或设计权衡。

风险与影响

风险：测试覆盖率可能降低，移除的冗余测试（如 TestEAGLERadixCache 中的多配置测试）或许捕捉边缘情况；合并类可能掩盖特定配置错误。影响：对用户无感知；系统 CI 时间减半，提升开发效率；团队需监控测试有效性以确保质量。

关联脉络

与近期 PR 如 #21554（移除冗余 PCG 测试）和 #21787（移除冗余 MoE 测试）一脉相承，反映团队持续优化 CI 测试套件以减少冗余、提升效率的趋势，属于测试维护和 CI 性能改进的常规工作。