PR #22652 完整报告

sgl-project/sglang

Simplify test_chunked_prefill; remove redundant tests

合并时间：2026-04-13 11:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22652

执行摘要

此 PR 简化了 chunked prefill 调度测试套件，通过移除冗余测试用例并重构剩余测试，旨在缩短 CI 执行时间约 50%（从 360 秒降至 180 秒）。核心变更包括删除原测试文件、创建新文件以专注于混合 chunk 配置覆盖，并引入严格内存检查以增强调试能力。

功能与动机

动机：减少 CI 时间，优化测试效率。PR body 指出："Remove redundant tests in test_chunked_prefill.py to reduce CI time"，具体移除三个测试：默认路径（被其他 MMLU 集成测试隐式覆盖）、禁用基数缓存路径（已在 test_no_overlap_scheduler.py 中覆盖）和多请求测试（缺乏准确性断言且很少触发 chunking）。

实现拆解

按模块拆解关键改动：

- 测试文件重构：
  - 移除 test/registered/scheduler/test_chunked_prefill.py，删除五个测试方法。
  - 新增 test/registered/scheduler/test_mixed_chunked_prefill.py，定义两个测试类：
- TestMixedChunkedPrefill：启用混合 chunk（--enable-mixed-chunk），设置 chunked_prefill_size=32，使用 GSM8K 混合评估准确性。
- TestMixedChunkedPrefillNoRadixCache：额外禁用基数缓存（--disable-radix-cache）。
- CI 配置更新：调整 register_cuda_ci 和 register_amd_ci 的 est_time 从 360 秒降至 180 秒。
- 测试环境强化：在 setUpClass 中启用严格内存检查（SGLANG_ENABLE_STRICT_MEM_CHECK_DURING_BUSY=2），提供更详细的日志输出。

评论区精华

review 讨论来自 gemini-code-assist[bot]，要点如下：

- chunked_prefill_size 设置：> "The test test_mixed_chunked_prefill does not explicitly set chunked_prefill_size" – 建议明确设置以正确触发 chunking 逻辑。
- 多请求测试覆盖：> "it is recommended to restore and adjust the multi-request test case" – 强调并发请求调度的重要性。 但作者未采纳这些建议，基于现有覆盖论证冗余，

PR 已合并且 CI 通过。

风险与影响

- 技术风险：
  - 测试覆盖率降低：移除的多请求测试可能未在其他地方覆盖并发场景，增加潜在 bug 漏检风险。
  - 严格内存检查可能引入误报：在高负载或边缘情况下，内存检查可能导致测试失败，需监控 CI 稳定性。
- 影响范围：
  - CI 时间减少，提升开发迭代速度。
  - 测试套件更简洁，便于维护，但团队需确保剩余测试（混合 chunk + 禁用基数缓存）充分覆盖关键路径。

关联脉络

此 PR 是测试优化系列的一部分：

- PR #22647：提取暂停 / 恢复测试工具包，重命名测试文件，与本 PR 的文件重构和调度测试主题直接相关。
- PR #20908 和 #21875：涉及调度修复和会话管理测试，反映团队持续关注调度可靠性和测试覆盖。 整体趋势显示仓库正通过移除冗余测试、增强关键场景覆盖来优化 CI 效率，同时平衡风险与维护成本。