PR #27303 完整报告

sgl-project/sglang

Use level-1 (quiet) busy memory check in chunked-prefill and streaming tests

合并时间：2026-06-05 05:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27303

执行摘要

- 一句话：测试中繁忙内存检查降级为安静模式
- 推荐动作：可直接合入，变更简单且意图明确。建议在日后 CI 调试时，若怀疑内存泄漏相关故障，可临时切换回 level 2 获取详细日志。

功能与动机

PR body 明确指出："Same per-step leak assertion, far less CI log spam." 目的是减少 CI 日志噪音，优化调试体验。

实现拆解

1. 修改 chunked-prefill 测试：在 test/registered/scheduler/test_mixed_chunked_prefill.py 中，将 setUpClass 里的环境变量覆盖值从 2 改为 1。
2. 修改 streaming-session 测试 fixture：在 python/sglang/test/server_fixtures/streaming_session_fixture.py 中，更新了类文档字符串和 setUpClass 中的环境变量覆盖值，同样从 2 改为 1。
3. 未修改其他逻辑：两个测试的启动参数、断言逻辑和 teardown 保持不变，仅切换了内存检查的日志详细级别。

关键文件：

- test/registered/scheduler/test_mixed_chunked_prefill.py（模块 测试；类别 test；类型 test-coverage）：chunked-prefill 测试中环境变量值从 2 改为 1，减少 CI 日志噪音。
- python/sglang/test/server_fixtures/streaming_session_fixture.py（模块 测试；类别 test；类型 test-coverage）：streaming session 测试 fixture 中环境变量值从 2 改为 1，同步更新文档注释。

关键符号：未识别

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。level 1 仍会按 step 检查内存泄漏，仅在泄漏时输出日志，不会影响泄漏检测的正确性。唯一可能的变化是调试时若没有泄漏则看不到池统计信息，但这正是预期的安静行为。
- 影响：影响范围仅限于 CI 日志输出：chunked-prefill 和 streaming-session 测试的日志将大幅减少，便于在大量测试输出中快速定位真正的问题。对用户功能无影响。
- 风险标记：测试变更

关联脉络

- PR #27238 Add quiet mode for busy mem check (level 1: buffer + dump on leak): 该 PR 实现了 level 1 安静模式，当前 PR 是应用该模式到测试中。