PR #21349 完整报告

sgl-project/sglang

[CI] Reduce session correctness test to 30 turns to fix flakiness

合并时间：2026-03-25 09:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21349

执行摘要

本 PR 将流式会话正确性测试的轮数从 150 减少到 30，以解决由 GPU 浮点非确定性引起的 CI 测试 flakiness，提高测试稳定性，同时保持其他延迟测试不变。

功能与动机

动机是解决测试 flakiness 问题。PR body 明确指出："GPU floating-point non-determinism causes greedy decoding to diverge on long multi-turn contexts, leading to cascading mismatches"，并引用了一个失败的 CI 运行。这表明在长上下文测试中，浮点计算的非确定性导致输出不匹配，因此需要调整测试参数来缓解此问题。

实现拆解

实现非常简单，仅修改了一个文件：test/registered/sessions/test_session_latency.py。关键改动如下：

- 在 test_streaming_session_correctness 函数中，引入局部变量 correctness_turns = 30。
- 将 _run_concurrent_session 调用中的 num_turns 参数设为该变量，例如：python reg = self._run_concurrent_session(streaming=False, num_concurrent=1, num_turns=correctness_turns)
- 其他测试函数如 test_regular_session 和 test_streaming_session 保持 150 轮不变，确保延迟测试不受影响。

评论区精华

该 PR 没有收到任何 review 评论，由作者直接合并，表明变更被认为直接且必要，无技术争议或设计权衡讨论。

风险与影响

风险：减少测试轮数可能降低对长上下文场景的验证强度，潜在隐藏浮点非确定性问题。具体到 test_streaming_session_correctness 函数，轮数减少后，可能无法充分捕获 GPU 计算在更长时间序列中的不匹配，从而遗漏潜在错误。 影响：正面影响是 CI 测试更稳定，减少不必要的失败和资源浪费；负面影响是测试覆盖略有降低，可能需要在未来通过其他方式（如增强测试设计或增加容忍度）来弥补。

关联脉络

与此 PR 相关的其他 CI 优化 PR 包括：

- PR 21305：增加缓存刷新超时，提升 CI 稳定性。
- PR 21330：默认启用 failfast，优化测试执行以减少不稳定测试的干扰。 这些 PR 共同反映了团队对 CI 测试可靠性的持续改进趋势，旨在减少 flakiness 和提高开发效率。