PR #21276 完整报告

sgl-project/sglang

Revert "fix: use consistent time denominator for throughput metrics in bench_one_batch_server"

合并时间：2026-03-24 13:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21276

PR 分析报告：回滚吞吐量度量计算

执行摘要

此 PR 回滚了 PR #19223 中对 bench_one_batch_server 测试吞吐量计算的更改，恢复使用总延迟作为分母，核心影响是测试度量的准确性，决策基于潜在问题回滚，属于常规维护操作。

功能与动机

PR 动机是撤销 PR #19223 的修复，该修复原本旨在使用一致的时间分母（last_ttft 和 latency-last_ttft）计算吞吐量。回滚原因未在 PR body 中详述，但 issue 评论提及讨论在 #18712，暗示原修复可能引入度量不一致或其他问题，因此决定回退到原始代码以保持稳定性。

实现拆解

仅修改了文件 python/sglang/test/bench_one_batch_server_internal.py 中的 run_one_case 函数。具体变更如下：

- 原始代码：python input_throughput = batch_size * input_len / last_ttft output_throughput = batch_size * output_len / (latency - last_ttft)
- 回滚后代码：python input_throughput = batch_size * input_len / latency output_throughput = batch_size * output_len / latency 这恢复了使用总延迟 latency 作为分母的计算方式，整体吞吐量公式 overall_throughput 保持不变。

评论区精华

无直接 review 讨论。issue 评论（来自 nvjullin）指出："Discussion is happening at #18712"，这表明回滚决策可能源于外部讨论，但 PR 本身未提供详细交锋，建议参考关联 issue 以获取更多上下文。

风险与影响

- 技术风险：回滚可能重新引入 PR #19223 试图解决的问题，例如吞吐量度量分母不一致，导致性能测试结果不准确，特别是在 vLLM 后端场景下，可能误导性能优化工作。
- 影响范围：直接影响限于内部测试脚本 bench_one_batch_server_internal.py，对用户或生产系统无直接影响，但可能影响团队对模型性能的评估和基准测试的可靠性。

关联脉络

- 直接关联：PR #19223（被回滚的 PR）直接关联，原 PR 修改了相同的测试文件以“修复”吞吐量度量计算，此回滚操作逆转了该变更。
- 潜在关联：issue #18712 可能包含相关讨论，涉及吞吐量度量的设计权衡或测试问题，建议查阅以理解回滚背后的技术决策和更大上下文。近期历史 PR 多为性能优化、bugfix 和 CI 改进，此 PR 属于测试维护的小范围变更，未显式关联其他功能演进。