PR #42455 完整报告

vllm-project/vllm

[CI] Fix `test_async_scheduling.py` flakiness

合并时间：2026-05-13 05:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42455

执行摘要

- 一句话：修复异步调度测试因 rank 排序波动导致的 flakiness
- 推荐动作：可立即合并。但建议后续跟踪测试稳定性，若仍有 flakiness 可考虑使用绝对容差，并处理 None rank 情况。

功能与动机

PR #41411 后，测试开始比较 prompt logprobs，导致 rank 值可能相差 1 或 2，但原有 a.rank == b.rank 严格相等使得测试在不相关排序变化时失败，成为 CI 中的超级 flaky 问题。PR body 明确引用 Buildkite 失败日志并指出根因。

实现拆解

1. 在 tests/v1/e2e/general/test_async_scheduling.py 的 _logprobs_match 函数中，将第 432 行 a.rank == b.rank 修改为 a.rank == pytest.approx(b.rank, rel=0.005)，允许 rank 在 0.5% 相对容差内匹配。
2. 仅改动一行，不影响其他逻辑。

关键文件：

- tests/v1/e2e/general/test_async_scheduling.py（模块 测试；类别 test；类型 test-coverage；符号 _logprobs_match）：唯一修改文件，核心 flakiness 修复在此。

关键符号：_logprobs_match

评论区精华

gemini-code-assist[bot] 指出 rel=0.005 对于小整数 rank 过于严格（例如 rank<200 时 0.5% 容差不足以允许偏差 1），建议改用绝对容差 abs=2。同时指出 b.rank 可能为 None，会导致 TypeError。但 PR 最终未采纳该建议，仅保留相对容差。khluu 和 yewentao256 均批准了 PR。

- rank 容差类型选择 (correctness): PR 保持相对容差，未采纳建议。

风险与影响

- 风险：低风险。变更仅涉及测试断言，不影响生产代码。但 gemini-code-assist[bot] 指出的 None rank 问题未处理，若测试中出现 None rank 仍会引发 TypeError。相对容差对小

rank 可能不够宽松，导致 flakiness 未完全消除。

- 影响：影响范围限于 test_async_scheduling.py 测试。修复后 CI 中该测试的 flakiness 应显著减少，但可能因容差不充分仍有偶发失败。
- 风险标记：测试逻辑变更，可能未完全修复 flakiness

关联脉络

- PR #41411 引入 prompt logprob 比较的 PR（推测）：PR body 指出该 PR 引入后测试开始 flaky，是本 PR 的根因。