PR #38576 完整报告

vllm-project/vllm

vLLM Benchmark Suite perf regression after PR#32723

合并时间：2026-03-31 13:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38576

执行摘要

本 PR 修复了由于 PR#32723 取消默认 temperature 设置导致的性能基准测试回归，通过在多个基准测试配置文件中添加 temperature: 0 参数，确保输出确定性以准确检测性能问题。变更仅影响 CI 测试套件，风险低，已直接合并。

功能与动机

PR#32723 移除了 temperature 的默认值 0，导致基准测试输出路径方差增加，影响性能回归的识别。PR body 明确表述："Since we need to identify perf regression issue so we need deterministic results." 因此，本 PR 旨在恢复确定性输出，通过固定 temperature 为 0 来消除方差干扰。

实现拆解

所有改动均位于 .buildkite/performance-benchmarks/tests/ 目录下的 JSON 配置文件，包括：

- serving-tests.json：通用基准测试配置，添加 "temperature": 0。
- serving-tests-arm64-cpu.json：针对 ARM64 CPU 后端。
- serving-tests-cpu.json、serving-tests-cpu-asr.json、serving-tests-cpu-text.json：针对不同 CPU 测试场景。
- serving-tests-hpu.json：针对 HPU 后端。

每个文件的变更类似，仅在相关配置对象中添加键值对，无代码逻辑调整。示例代码片段：`{ "temperature":0, "num_prompts":200 }`

评论区精华

Review 中无技术讨论。仅有自动化 bot 评论指出无反馈，以及合并者 njhill 的简短批准（"Thanks @louie-tsai!"），表明变更被认可且无争议。

风险与影响

风险较低：

- 仅配置变更，不触及核心代码，无回归或安全风险。
- 但提交历史显示第二个提交补丁了 ARM CPU、NV GPU 和 Gaudi，文件列表未完全覆盖，可能遗漏部分测试文件，导致性能检测不准确。

影响有限：

- 直接作用于 CI 性能基准测试，提升回归检测可靠性。
- 对最终用户和系统运行时无影响。

关联脉络

本 PR 直接关联 PR#32723（未在提供历史列表中），修复其引起的基准测试不确定性。同仓库近期历史 PR 多涉及 bugfix、性能优化和重构（如 #38546 清理 KVConnector、#36847 新增推测解码），但本 PR 更侧重于 CI 基础设施维护，反映团队对测试稳定性的重视。