

PR #46568 完整报告

vllm-project/vllm

Upgrade tpu-inference to v0.23.0

合并时间: 2026-06-24 12:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46568>

执行摘要

将 TPU 推理依赖库 `tpu-inference` 从 v0.22.1 升级到 v0.23.0, 仅修改 `requirements/tpu.txt` 中的一个版本号, 已通过 TPU CI 验证。

功能与动机

升级到最新稳定版本以获取 bug 修复和潜在改进。PR body 声明已验证于 TPU 推理 CI。

实现拆解

1. 变更入口: `requirements/tpu.txt`, 依赖声明文件。
2. 具体变更: 将 `tpu-inference==0.22.1` 替换为 `tpu-inference==0.23.0`。
3. 测试验证: PR 作者称已在 `tpu-inference` CI 上验证成功, 但未提供更多细节或测试结果截图。

无。

评论区精华

无 review 讨论。

风险与影响

- 风险: 极低。但新版本可能引入不兼容变更, 需关注后续 TPU 相关测试是否稳定。
- 影响: 仅限于 TPU 推理用户, 其他后端不受影响。

关联脉络

无关联 Issue 或历史 PR。