PR #37923 完整报告

vllm-project/vllm

[Bugfix] Force continuous usage stats when CLI override is enabled

合并时间：2026-03-25 01:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37923

执行摘要

该 PR 修复了 --enable-force-include-usage CLI 标志未正确启用连续使用统计的 bug，通过修改 should_include_usage() 函数逻辑并添加回归测试，确保服务器级覆盖始终包括连续使用统计，影响范围仅限于前端使用统计功能。

功能与动机

动机是确保当服务器启用 --enable-force-include-usage 标志时，响应中始终包含最终使用统计和连续使用统计，修复了原有逻辑中连续使用统计可能被 stream options 禁用的错误。PR body 明确指出“make --enable-force-include-usage return (True, True) from should_include_usage()”以解决此问题，无需外部 issue 驱动。

实现拆解

- 核心逻辑变更：在 vllm/entrypoints/utils.py 中，should_include_usage() 函数现在当 enable_force_include_usage 为 True 时直接返回 (True, True)，简化了条件判断。python if enable_force_include_usage: return True, True
- 测试覆盖：在 tests/entrypoints/test_utils.py 中添加了参数化测试 test_should_include_usage_force_enables_continuous_usage，验证在不同 stream options（如 None、include_usage=False 等）下强制覆盖的行为。
- 测试更新：更新了 tests/entrypoints/openai/chat_completion/test_enable_force_include_usage.py 中的断言，移除 assert chunk.usage is None 以匹配新逻辑，确保测试通过。

评论区精华

Review 讨论较少，仅 gemini-code-assist[bot] 确认“The changes are sound and effectively resolve the issue.”，simon-mo 批准无评论。无技术争议或深度讨论，变更被快速接受。

风险与影响

- 风险：变更范围小，风险低。主要风险是修改了核心函数 should_include_usage()，如果 early return 逻辑错误，可能影响其他 stream options 处理，但测试覆盖了多种场景，降低了回归风险。

- 影响：仅影响使用 *--enable-force-include-usage* 标志的用户，确保连续使用统计正确启用，对系统其他部分无影响，属于前端行为修复。

关联脉络

PR body 提及 #24477 添加 token-details 字段到连续使用统计，但本 PR 专注于 CLI 覆盖语义变更，无直接技术关联。在同仓库近期历史 PR 中，类似 bugfix PR 如 #37911 修复前端警告，但未修改相同文件，无其他跨 PR 关联脉络。