PR #37460 完整报告

vllm-project/vllm

[Core][Metrics][BugFix] Replace num_cached_tokens/num_external_computed_tokens with PrefillStats

合并时间：2026-04-14 16:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37460

执行摘要

本 PR 通过引入 PrefillStats 数据结构替换了原有的 num_cached_tokens 和 num_external_computed_tokens 字段，解决了调度器在 preemption 和 KV 传输失败下 prefill 统计不准确的问题，提升了 metrics 报告的正确性，并简化了代码路径，为后续清理奠定基础。

功能与动机

动机源于 PR #36859 中讨论的 Counters can only be incremented by non-negative amounts 错误，该错误由 num_cached_tokens 和 num_external_computed_tokens 在 preemption 下的不一致设置导致。PR body 进一步指出，随着 #38096 的合并，num_cached_tokens 不再用于 KV 传输失败恢复，因此本 PR 移除了脆弱的错误处理逻辑，改用 PrefillStats 在请求首次调度时记录统计信息，确保 metrics 的清晰性和准确性。

实现拆解

关键改动按模块分解：

- metrics 模块（vllm/v1/metrics/stats.py）：新增 PrefillStats dataclass，包含 num_prompt_tokens、num_local_cached_tokens 等字段，并修改 PromptTokenStats.update_from_output() 以使用新结构。
- scheduler 模块（vllm/v1/core/sched/scheduler.py）：在 schedule() 方法中首次调度 prefill 时设置 request.prefill_stats，并通过 SchedulerOutput 传递。
- engine 模块（vllm/v1/engine/__init__.py）：更新 EngineCoreOutput，移除旧字段，添加 prefill_stats 可选字段。
- core 模块（vllm/v1/request.py）：在 Request 类中新增 prefill_stats 字段和 take_prefill_stats() 方法，移除 num_cached_tokens 和 num_external_computed_tokens。
- 测试更新：同步修改了多个测试文件，如 tests/v1/metrics/test_stats.py，以验证新逻辑的正确性。

评论区精华

Review 讨论中突出以下交锋：

1. 双计数争议：gemini-code-assist[bot] 指出 metrics 计算可能存在双计数，但 markmc 解释这是预期行为，并引用 #33289 说明设计意图，结论是计划在 #38709 中进一步处理。

2. 设计简化：orozery 质疑 set_once() 的必要性，认为调度器应直接负责统计，经讨论后第二个 commit 移除了该行为，代码更简洁。
3. 测试边界：gemini-code-assist[bot] 提到测试断言在提示长度整除 BLOCK_SIZE 时可能不准确，markmc 回应未来场景需调整，但当前暂不处理。

风险与影响

- 风险：核心调度路径变更可能引入回归，尤其是在复杂 preemption 场景；移除旧字段可能破坏依赖代码；测试覆盖不足可能隐藏边缘情况 bug。
- 影响：用户将获得更准确的 prompt tokens metrics，但需更新监控工具；系统代码更健壮，减少了错误处理复杂度；团队需适应新结构，后续 PR 可能继续清理。

关联脉络

本 PR 与多个历史 PR 紧密相关：

- 36859 直接启发了本 PR，修复了字段设置不一致问题。
- 38096 为移除 KV 传输失败恢复逻辑提供了前提。
- 38709 计划解决 review 中提及的 metrics 双计数问题，显示功能演进的连续性。 整体上，这一系列变更反映了 vLLM 在核心调度和 metrics 模块上持续优化，以提升可靠性和可维护性。