# PR #39693 完整报告

vllm-project/vllm

[Core][Metrics] Remove unused `SchedulerStats.encoder_cache_usage`

合并时间：2026-04-15 00:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39693

## 执行摘要

- 一句话：移除调度器统计中未使用的编码器缓存使用率字段，清理无用代码。
- 推荐动作：该 PR 变更简单直接，适合快速浏览以了解代码清理决策。值得关注的是团队对未使用代码的处理原则：优先移除而非保留，强调指标应面向用户设计。

## 功能与动机

PR #33452 添加了 *encoder_cache_usage* 字段，但该指标未在任何地方向用户公开。根据 PR body 和 Issue 评论，作者 markmc 认为不应无限期保留可能未被使用的代码，特别是来自外部贡献者（如 Meta）的指标。他指出，如果该指标对用户有价值，可以轻松重新添加并集成到 Prometheus 中。

## 实现拆解

该 PR 删除了两个文件中的相关代码：

1. vllm/v1/*core/sched/scheduler.py*：移除了 *make_stats* 方法中 *encoder_cache_usage* 的赋值，并删除了整个 *_get_encoder_cache_usage* 方法（共 9 行）。
2. vllm/v1/*metrics/stats.py*：从 SchedulerStats 数据类中移除了 *encoder_cache_usage* 字段定义（1 行）。

关键文件：

- vllm/v1/*core/sched/scheduler.py*（模块 *core/scheduler*）：移除了 *encoder_cache_usage* 的计算逻辑和 *_get_encoder_cache_usage* 方法，是核心调度器模块的清理。
- vllm/v1/*metrics/stats.py*（模块 *metrics*）：从 SchedulerStats 数据类中移除了未使用的 *encoder_cache_usage* 字段定义。

关键符号：*make_stats*, *_get_encoder_cache_usage*

## 评论区精华

Review 中讨论较少，*gemini-code-assist[bot]* 确认了变更内容，robertgshaw2-redhat 直接批准。主要讨论在关联 Issue 中：DarkLight1337 建议联系 Meta 确认是否仍需该字段，markmc 回应不应仅因可能被某个分支使用而无限期保留代码，强调若指标有价值可重新添加并集成到 Prometheus。

- 是否应保留未使用的指标字段 (design): 决定移除字段，若未来有需求可重新添加并集成到 Prometheus。

## 风险与影响

- 风险：风险较低：
  1. 该字段从未向用户公开，移除不会影响现有功能或 API。
  2. 删除的是未使用的代码，不会引入回归问题。
  3. 若未来需要该指标，需重新实现并测试，但根据 PR body 描述可轻松完成。
- 影响：影响范围有限：
  1. 对用户无影响，因为该字段未在用户可见的指标中暴露。
  2. 减少代码复杂度和维护负担，提升代码清晰度。
  3. 为未来指标设计提供空间，可重新添加更完善的 Prometheus 集成。
- 风险标记：低风险变更，清理未使用代码

## 关联脉络

- PR #33452 未知：PR body 提到该字段由 PR #33452 添加，是本 PR 清理的源头。
- PR #37460 [Core][Metrics][BugFix] Replace num_cached_tokens/num_external_computed_tokens with PrefillStats: 同属 *core* 和 metrics 标签的 PR，涉及调度器统计和指标重构，可对比指标清理与重构模式。