PR #37498 完整报告

vllm-project/vllm

[Frontend][Responses API] Fix arrival_time recording for TTFT on initial request

合并时间：2026-03-23 17:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37498

执行摘要

- 一句话：修复 responses API 中 arrival_time 记录错误，以准确测量 TTFT。
- 推荐动作：对于负责性能度量或 API 实现的工程师，建议精读此 PR 以理解 arrival_time 定义的重要性和当前修复。同时，关注 markmc 指出的其他问题，可能需要在后续 PR 中解决。

功能与动机

PR body 指出，在 responses API 中，arrival_time 应该在 tokenization 之前定义以准确测量 TTFT，但当前 GPT-OSS 实现中，arrival_time 从 input_processor.py 获取，发生在 tokenization 之后，这导致性能度量偏差。issue 评论中 DarkLight1337 也提到，之前定义在 tokenization 之后很奇怪，已改为之前以获取更准确的 TTFT 测量。

实现拆解

主要改动包括：1) 在 vllm/entrypoints/openai/responses/serving.py 的 _make_request_with_harmony 函数中，添加 arrival_time = time.time() 并在 tokenization 前赋值给 engine_prompt["arrival_time"]；2) 更新 docs/design/metrics.md 文档，说明 arrival_time 当前定义为 tokenization 开始时，并添加注释指出这是一个非平凡话题，引用相关讨论。

关键文件：

- vllm/entrypoints/openai/responses/serving.py（模块 openai responses API）：核心修复代码，在 Harmony 流中正确记录 arrival_time 以修复 TTFT 度量
- docs/design/metrics.md（模块 文档）：更新文档，说明 arrival_time 定义和背景，帮助用户理解 metrics 含义

关键符号：_make_request_with_harmony

评论区精华

review 中，gemini-code-assist[bot] 确认实现正确。markmc 指出还有更严重的问题，涉及多轮工具调用中 arrival_time 记录在工具完成之后而不是 tokenization 之前，建议单独修复。issue 评论中，DarkLight1337 和 qandrew 讨论了 arrival_time 定义的历史和正确性，结论是应该定义在 tokenization 之前以准确测量 TTFT。

- arrival_time 定义正确性 (correctness): 已修复为 tokenization 之前，达成共识

- 其他未解决的 arrival_time 问题 (correctness): 未解决，需要后续 PR 处理

风险与影响

- 风险：风险较低，因为代码修改范围小且直接，但 markmc 提到的未解决问题（多轮工具调用中的 arrival_time 错误）可能影响部分场景的性能度量准确性。文档更新可能不完整，metrics 定义仍需进一步澄清，存在理解偏差风险。
- 影响：修复后，TTFT 度量将更准确，直接影响 API 用户获取可靠的性能数据，有助于监控和基准测试。对系统功能无直接影响，但提升了性能评估的精确性。团队需注意后续修复多轮工具调用问题。
- 风险标记：未解决多轮工具调用问题，文档定义不完整

关联脉络

- 暂无明显关联 PR