

PR #47815 完整报告

vllm-project/vllm

[Bugfix][OpenAI] Fix streamed completion logprob offsets with echo

合并时间: 2026-08-27 18:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47815>

执行摘要

- 一句话: 修复流式 echo 日志偏移, 改为按 `delta_text` 推进
- 推荐动作: 该 PR 值得快速阅读, 尤其是理解 `delta_text` 与 `output.text` 在 echo 场景下的语义差异。一行修复背后涉及流式偏移状态的正确性, 作者通过详尽的 E2E 验证提升了可信度。建议后续补充自动化回归测试, 将此类边界条件固化。

功能与动机

在 OpenAI 兼容接口中, 当用户同时开启 `stream=True`、`echo=True` 和 `logprobs` 时, 返回的 `logprobs.text_offset` 是静默错误的。PR body 明确指出: 第一个流块可能同时包含回显的 prompt 和生成文本, 但流状态仅按生成文本长度推进, 导致后续每个块的 `text_offset` 都左移了 prompt 长度。作者还说明了此 PR 与现有修复的区分: 无需重复 #44106 (V2 prompt logprobs)、#44938 (Rust prompt-only)、#19379/#19380 (`return_tokens_as_token_ids`) 等。

实现拆解

该 PR 仅修改一个文件的一行代码, 涉及 OpenAI 兼容接口的流式补全逻辑。

1. 变更入口: 修改 `vllm/entrypoints/openai/completion/serving.py` 中的 `completion_stream_generator` 函数。
2. 核心变更: 将 `previous_text_lens[i] += len(output.text)` 改为 `previous_text_lens[i] += len(delta_text)`, 从而按实际发出的文本长度推进文本偏移状态。
3. 原因: 在 `echo=True` 时, 第一个块中 `delta_text` 包含回显的 prompt + 生成文本, 而 `output.text` 仅包含生成文本, 旧实现导致后续 `text_offset` 偏小; 使用 `delta_text` 后, 偏移与实际发出的文本一致。
4. 影响范围: 非 echo 场景下 `delta_text == output.text`, 行为不变; `max_tokens=0` 路径为终止块, 状态推进无下游影响。
5. 测试配套: PR 未新增自动化测试, 但作者提供了详细的服务器端 E2E 复现脚本和验证结果。

关键文件:

- `vllm/entrypoints/openai/completion/serving.py` (模块 前端; 类别 `source`; 类型 `core-logic`; 符号 `completion_stream_generator`): 核心修复文件, 修改流式补全中文本偏移状态的推进逻辑。

关键符号: completion_stream_generator

关键源码片段

vllm/entrypoints/openai/completion/serving.py

核心修复文件，修改流式补全中文本偏移状态的推进逻辑。

```
# vllm/entrypoints/openai/completion/serving.py
# 在流式补全生成器中，推进文本偏移状态。
# 修复前使用 output.text 的长度，在 echo=True 时第一个块会包含 prompt，导致后续偏移错误。
# 修复后改用 delta_text（实际发出的文本），确保偏移正确。
async def completion_stream_generator(...):
    # ... 省略上下文 ...
    if request.logprobs is not None:
        assert out_logprobs is not None, "Did not output logprobs"
        logprobs = self._create_completion_logprobs(
            token_ids=delta_token_ids,
            top_logprobs=out_logprobs,
            num_output_top_logprobs=request.logprobs,
            tokenizer=tokenizer,
            logprob_token_ids=request.logprob_token_ids,
            initial_text_offset=previous_text_lens[i], # 使用当前偏移
            return_as_token_id=request.return_tokens_as_token_ids,
        )
    else:
        logprobs = None

    # 关键修复：按实际发出的文本长度推进偏移，而非 output.text。
    # echo=True 时，delta_text 包含 prompt 和生成文本，output.text 仅含生成文本。
    previous_text_lens[i] += len(delta_text)
    previous_num_tokens[i] += len(output.token_ids)
    # ... 后续逻辑 ...
```

评论区精华

主要审核意见来自 NolenLiang 的批准评论：

The state update now matches the text actually emitted by the stream: on the first `echo=True` chunk, `delta_text` includes the prompt while `output.text` does not. Subsequent `initial_text_offset` values therefore advance correctly, and the non-echo path is unchanged because `delta_text == output.text` there.

A committed streaming+echo offset regression test would still be useful, but I do not consider that a correctness blocker for this narrowly scoped fix.

核心讨论聚焦于修复的合理性和测试覆盖的缺失。评审者认可一行改动的正确性（基于对 `delta_text` 与 `output.text` 语义的清晰解释），但建议补充流式 +echo 偏移的回归测试。没有未解决的疑虑。

- 修复正确性与测试覆盖 (testing): 评审者批准，测试缺失不阻塞合入。

风险与影响

- 风险：风险极低，因为：
 - 代码变更仅一行，逻辑简单，非核心路径。
 - 非 echo 场景行为完全不变。
 - 作者提供了真实服务器的 E2E 验证，包括修复后、反转后和恢复后的对比。
 - 潜在风险：缺少自动化回归测试，未来可能再次引入同类回归。此外，若 `delta_text` 与 `output.text` 在极端情况下不一致（例如某些特殊 token 处理），偏移可能不符合预期，但当前逻辑中二者在非 echo 时完全一致。
 - 影响：影响范围较小，仅影响 OpenAI 兼容接口中同时使用 `stream=True`、`echo=True` 和 `logprobs` 的场景。修复后，这些用户的 `text_offset` 将正确反映文本位置，提升数据准确性。对系统性能和稳定性无影响。团队方面，该修复避免了潜在的 API 数据误导问题，但缺少测试可能增加维护成本。
- 风险标记：缺少测试覆盖

关联脉络

- PR #44106 V2 prompt logprobs 相关：本 PR 说明不覆盖 V2 prompt logprobs，区分范围。
- PR #44938 Rust prompt-only completions: 本 PR 说明不覆盖 Rust prompt-only，区分范围。
- PR #19379 `return_tokens_as_token_ids` 相关：本 PR 说明不覆盖 `return_tokens_as_token_ids`，区分范围。
- PR #19380 `return_tokens_as_token_ids` 相关：本 PR 说明不覆盖 `return_tokens_as_token_ids`，区分范围。