

PR #44461 完整报告

vllm-project/vllm

[Bugfix][Voxtral Realtime] Fix token feedback timeout silent hang

合并时间: 2026-07-05 20:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44461>

执行摘要

- 一句话: 移除 feed_tokens 超时, 修复 Voxtral 静默挂起
- 推荐动作: 值得合并。修复逻辑清晰, 改动极小, 且经 E2E 测试验证。建议后续考虑为队列等待添加更长的超时或健康检查机制, 但当前方案已满足需求。

功能与动机

修复 #36015 (连续流式音频约 10 分钟后无输出但连接不断) 和 #35863 (第三并发会话无转录输出)。根本原因为 feed_tokens 中 `asyncio.wait_for` 将正常的空闲等待视为异常, 导致后台任务退出, Transcript 停滞。

实现拆解

1. 移除超时相关的导入: 删除文件顶部的 `from vllm.envs import VLLM_ENGINE_ITERATION_TIMEOUT_S` 导入。
2. 简化 feed_tokens 协程: 将 `all_outputs = await asyncio.wait_for(input_stream.get(), timeout=VLLM_ENGINE_ITERATION_TIMEOUT_S)` 改为 `all_outputs = await input_stream.get()`, 去掉超时包装。
3. 更新注释: 将“Feed output tokens back into buffer in background”改为强调空闲等待是正常的, 任务生命周期由 finally 块管理。
4. 未修改其他部分: feed_audio、主循环、任务取消逻辑均保持不变。

关键文件:

- `vllm/model_executor/models/voxtral_realtime.py` (模块 模型执行器; 类别 source; 类型 core-logic; 符号 feed_tokens) : 核心修复文件: 移除了 feed_tokens 协程中针对 `input_stream.get()` 的超时, 解决了 Voxtral Realtime 静默挂起的根本原因。

关键符号: `feed_tokens`

关键源码片段

`vllm/model_executor/models/voxtral_realtime.py`

核心修复文件: 移除了 feed_tokens 协程中针对 `input_stream.get()` 的超时, 解决了 Voxtral Realtime 静默挂起的根本原因。

```
# vllm/model_executor/models/voxtral_realtime.py
```

```
# 删除导入: from vllm.envs import VLLM_ENGINE_ITERATION_TIMEOUT_S

# ... 其他代码 ...

# Feed output tokens back into the buffer. Idle waits are normal here;
# request cleanup still cancels this task through the finally block.
async def feed_tokens():
    while True:
        # 直接调用 input_stream.get(), 不再使用 asyncio.wait_for 超时
        # 之前: all_outputs = await asyncio.wait_for(input_stream.get(), timeout=...)
        all_outputs = await input_stream.get()
        await buffer.append_tokens(all_outputs[-1:])
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。移除超时后，若 `input_stream` 在某次 `get()` 后不再有新项目，`feed_tokens` 将永远阻塞，但任务通过 `finally` 块中的 `token_task.cancel()` 取消；请求生命周期结束、客户端断开、引擎异常等都会导致生成器结束，从而触发取消。仅当请求已结束但 `cancel()` 未生效（如事件循环异常）时可能出现僵尸任务，但此场景原代码也无法处理。
- 影响：正面影响：消除 Voxtral Realtime 在单会话空闲或多会话并发时的静默挂起，提升服务稳定性。影响范围仅限于 Voxtral Realtime 模型。无性能或兼容性影响。
- 风险标记：低风险

关联脉络

- PR #36089 [Bugfix][Voxtral Realtime] Retry `feed_tokens` on timeout: 之前的修复尝试，保留了超时并在超时后重试；本 PR 采用更彻底的方案，移除超时本身。
- PR #37483 [Bugfix] Fix Voxtral Realtime test timeout and connection validation: 仅修改了测试 timeout 和连接验证，未触及 `feed_tokens` 生产代码。