PR #50451 完整报告

vllm-project/vllm

[CI] Fix `tests/entrypoints/multimodal/openai/chat_completion/test_audio.py::test_chat_streaming_audio`

合并时间：2026-07-31 07:18

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50451

执行摘要

- 一句话：修复音频流测试 flaky 问题
- 推荐动作：建议合并。该 PR 以最小改动解决了 CI flaky 问题，但建议未来考虑更根本的修复（如放宽 token 级别匹配为语义匹配）或调查 prefix caching 对 ultravox 模型数值结果的影响程度。对于阅读者，可作为测试稳定性修复的参考案例。

功能与动机

关联 Issue #50385 报告 test_chat_streaming_audio 测试在 CI 中 flaky 失败，streaming 和 non-streaming 请求在 temperature=0 时返回不同 token 序列（如 'A short title' vs 'A possible short title'）。PR body 说明 prefix caching hit 可能干扰数值一致性，因此全局禁用 prefix caching 以修复该测试。

实现拆解

1. 移除平台条件判断：在 tests/entrypoints/multimodal/openai/chat_completion/test_audio.py 中，删除导入 from vllm.platforms import current_platform 及依赖该函数的 _ROCM_ARGS 变量。
2. 全局应用 --no-enable-prefix-caching：将 server fixture 中的 *_ROCM_ARGS 替换为硬编码的 "--no-enable-prefix-caching"，使得所有平台上都禁用 prefix caching，而非仅 ROCm。
3. 清理冗余注释：移除原仅针对 ROCm 的注释，减少代码噪音。

关键文件：

- tests/entrypoints/multimodal/openai/chat_completion/test_audio.py（模块 测试；类别 test；类型 test-coverage）：该文件是唯一变更的文件，修改了 server fixture 的参数，将 --no-enable-prefix-caching 从仅 ROCm 改为全局启用，并清理了相关导入和变量。

关键符号：未识别

关键源码片段

tests/entrypoints/multimodal/openai/chat_completion/test_audio.py

该文件是唯一变更的文件，修改了 server fixture 的参数，将 --no-enable-prefix-caching 从仅 ROCm 改为全局启用，并清理了相关导入和变量。

```
# tests/entrypoints/multimodal/openai/chat_completion/test_audio.py
# 变更前后对比：移除了 ROCm 条件判断，统一禁用 prefix caching

# 删除的导入行：
# from vllm.platforms import current_platform

# 删除的变量与注释：
# # Disable prefix caching on ROCm to reduce non-determinism in
# # streaming-vs-non-streaming comparisons.
# _ROCM_ARGS = ["--no-enable-prefix-caching"] if current_platform.is_rocm() else []

# server fixture 中的参数列表变更：
@pytest.fixture(scope="module")
def server():
    args = [
        "--dtype",
        "float32",
        "--max-model-len",
        "2048",
        "--max-num-seqs",
        "5",
        "--enforce-eager",
        "--trust-remote-code",
        "--limit-mm-per-prompt",
        json.dumps({"audio": MAXIMUM_AUDIOS}),
        "--no-enable-prefix-caching", # 之前是 *_ROCM_ARGS，现硬编码为全局启用
    ]

    with RemoteOpenAIServer(MODEL_NAME, args) as remote_server:
        yield remote_server
```

评论区精华

在 Issue #50385 的评论中，njhill 指出该测试之前并未失败，现在似乎稳定失败，建议理解什么发生了变化。NickLucche 回复认为 prefix caching 命中可能干扰数值，但测试本身 token 级别的匹配可能过于严格，可以适当放宽。最终决定通过全局禁用 prefix caching 修复，但未进一步调查根本原因或放宽断言精度。

- 测试 flaky 的根本原因分析 (question): NickLucche 认为 prefix caching hit 可能干扰数值，但未进行深入调查。最终通过禁用 prefix caching 修复，但未确认根本原因。

风险与影响

- 风险：该 PR 风险很低，仅修改测试配置参数，不影响任何生产代码。可能的风险是：如果其他测试或场景依赖该测试中 prefix caching 的默认开启状态，可能漏测相关行为。但测试本身的目的是验证流式与非流式一致性，禁用 prefix caching 在该上下文中合理。

- 影响：影响范围：仅影响单个测试文件 test_audio.py 中的一个测试函数 test_chat_streaming_audio。该测试在 CI 中的 Entrypoints Integration (Multimodal) 步骤中运行。 影响程度：消除该测试的 flaky 失败，使得 CI 更稳定。对其他测试无影响。
- 风险标记：测试稳定性修复

关联脉络

- 暂无明显关联 PR