

PR #46823 完整报告

vllm-project/vllm

[ROCm] [CI] fix transcription flakiness AMD: Entrypoints Integration (API Server OpenAI - Part 1) (mi325_1)

合并时间: 2026-06-26 21:33

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46823>

执行摘要

- 一句话: 修复 AMD CI 音频转录测试的随机失败
- 推荐动作: 可直接合并。此 PR 虽小, 但体现了在 CI 稳定性维护中通过明确测试参数来消除不确定性的良好实践。

功能与动机

AMD CI 的 Entrypoints Integration (API Server OpenAI - Part 1) 测试在 Buildkite 上出现不稳定失败 (<https://buildkite.com/vllm/ci/builds/74553>), 作者推测是语言自动检测失败导致, 因此显式指定音频语言为英语。

实现拆解

在 `tests/entrypoints/openai/test_run_batch.py` 中, 为 `INPUT_TRANSCRIPTION_HTTP_BATCH` 请求体添加 `"language": "en"` 字段, 确保 Whisper 模型在处理该测试音频时明确知道语言为英语, 避免自动检测失败。

关键文件:

- `tests/entrypoints/openai/test_run_batch.py` (模块 测试; 类别 test; 类型 test-coverage)
: 唯一变更文件, 通过添加 `language` 参数修复转录测试的随机失败问题。

关键符号: 未识别

关键源码片段

`tests/entrypoints/openai/test_run_batch.py`

唯一变更文件, 通过添加 `language` 参数修复转录测试的随机失败问题。

```
# 在 INPUT_TRANSCRIPTION_HTTP_BATCH 请求体中添加 "language": "en"
# 避免 Whisper 语言自动检测失败导致的测试不稳定
INPUT_TRANSCRIPTION_HTTP_BATCH = (
    json.dumps(
        {
            "custom_id": "request-1",
            "method": "POST",
            "url": "/v1/audio/transcriptions",
```

```
        "body": {
            "model": SPEECH_LARGE_MODEL_NAME,
            "file_url": AudioAsset("mary_had_lamb").url,
            "language": "en", # 显式指定语言为英语
            "response_format": "json",
        },
    }
)
+ "\n"
)
```

评论区精华

仅有一条 PR comment 来自作者“Let me check again.”，无 review 讨论。审核人 DarkLight1337 直接批准。

- 语言自动检测失败 (correctness): 通过添加 language 参数解决。

风险与影响

- 风险：风险极低。仅修改测试用例的请求参数，不涉及任何生产代码逻辑，不会影响系统行为。但需注意该测试依赖的音频文件 mary_had_lamb 本身是否为英语内容，若未来使用非英语音频可能需要调整。
- 影响：影响范围限定在 AMD CI 的转录测试，预期将减少该测试的随机失败，提高 CI 稳定性。对其他平台（如 NVIDIA）无影响。
- 风险标记：测试覆盖调整

关联脉络

- 暂无明显关联 PR