

PR #49853 完整报告

vllm-project/vllm

[CI] Fix speech correctness check rejecting improved WER

合并时间: 2026-07-26 14:24

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49853>

执行摘要

- 一句话: 修复 CI 中语音正确性测试的 WER 断言
- 推荐动作: 建议合并。该 PR 修复了测试断言语义, 逻辑清晰, 代码简洁, 风险很低。

功能与动机

在 Buildkite build #80224 中, Cohere 模型的 WER 为 11.698929, 低于基线 11.92, 但因 `torch.testing.assert_close` 的双向对称检查而失败。该 PR 旨在修复这一断言语义问题, 使 CI 能够正确接受改进的 WER。

实现拆解

1. 在 `test_transcription_api_correctness.py` 的 `test_wer_correctness` 函数中, 将 `torch.testing.assert_close(wer, expected_wer, atol=1e-1, rtol=1e-2)` 替换为单向上限比较。
2. 计算最大允许 WER: `max_wer = expected_wer + wer_atol + wer_rtol * abs(expected_wer)`。
3. 使用 `assert wer <= max_wer` 进行断言, 失败时输出清晰的错误信息。
4. 保留与之前相同的容忍度和上限边界, 确保仅改变方向。
5. 有意识不修改长格式测试 (`test_long_audio_wer_correctness`), 因为未观察到失败。

关键文件:

- `tests/entrypoints/speech_to_text/correctness/test_transcription_api_correctness.py` (模块 语音测试; 类别 test; 类型 test-coverage): 这是唯一变更的文件。将短格式 WER 断言从对称 `assert_close` 改为单向上限 `assert wer <= max_wer`。

关键符号: 未识别

关键源码片段

`tests/entrypoints/speech_to_text/correctness/test_transcription_api_correctness.py`

这是唯一变更的文件。将短格式 WER 断言从对称 `assert_close` 改为单向上限 `assert wer <= max_wer`。

```
# tests/entrypoints/speech_to_text/correctness/test_transcription_api_correctness.py
# 变更后：单向上限检查，避免因 WER 低于基线而误报
if expected_wer:
    wer_atol, wer_rtol = 1e-1, 1e-2
    # 按原始双向断言的公式计算最大允许 WER，保持上限一致
    max_wer = expected_wer + wer_atol + wer_rtol * abs(expected_wer)
    assert wer <= max_wer, (
        f"WER {wer:.6f} exceeds maximum allowed {max_wer:.6f} "
        f"(baseline {expected_wer:.6f})"
    )
```

评论区精华

无 review 讨论。PR 由维护者直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅修改一个测试函数中的断言逻辑，计算方式与之前的上限保持一致。不会影响生产代码或其它测试。唯一风险是如果长格式测试未来出现类似问题，需要单独处理。
- 影响：仅影响短格式语音正确性 CI 测试。不再因 WER 优于基线而失败，使 CI 更可靠。对用户无影响，对系统无影响。
- 风险标记：暂无

关联脉络

- PR #49773 [CI] Compute speech WER directly with jiwer: 该 PR 修复了 evaluate.load 依赖失败问题，但未改变断言方向；本 PR 是其中 assertion 语义问题的后续修复。