

PR #49773 完整报告

vllm-project/vllm

[CI] Compute speech WER directly with jiwer

合并时间: 2026-07-26 08:51

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49773>

执行摘要

- 一句话: 用 `jiwer` 替换 `evaluate` 修复 WER 测试崩溃
- 推荐动作: 这是一个直接且低风险的 bug 修复, 值得快速合并。它解决了上游依赖不兼容导致的 CI 中断, 且经过充分验证。无需精读, 但值得关注如何通过直接使用底层库来消除间接依赖。

功能与动机

修复 Issue #49771: Speech-to-text CI 作业中的所有三个 WER 正确性测试都因 `evaluate==0.4.3` 调用 `huggingface_hub.hf_api.HfFolder` (已在 `huggingface_hub==1.22.0` 中移除) 而失败。PR body 指出测试仅需要 WER, 而 `jiwer` 已是测试依赖, 因此直接使用它更简洁且无运行时在线依赖。

实现拆解

1. 替换导入: 在 `tests/entrypoints/speech_to_text/correctness/test_transcription_api_correctness.py` 中将 `from evaluate import load` 替换为 `from jiwer import wer`。
2. 简化 WER 计算: 在 `run_evaluation` 和 `run_longform_evaluation` 两个函数中, 将原来的两行代码 (先 `wer = load('wer')` 再 `wer.compute(...)`) 替换为单行 `wer_score = 100 * wer(references, predictions)`, 直接使用 `jiwer.wer` 函数。
3. 行为保持: 每个样本的对齐和全局错误聚合逻辑不变, `jiwer.wer` 的默认行为与 Hugging Face 的 WER 指标完全兼容 (已在 1900 个随机批次上验证)。
4. 测试验证: 通过 `pytest` 对修改后的测试文件进行 lint 检查, 并已验证 `jiwer` 的结果与 Hugging Face 聚合结果一致。完整的语音正确性套件由 Buildkite CI 覆盖。

关键文件:

- `tests/entrypoints/speech_to_text/correctness/test_transcription_api_correctness.py` (模块测试; 类别 `test`; 类型 `test-coverage`): 这是唯一变更的文件。它替换了 `evaluate.load` 调用, 改用 `jiwer.wer` 直接计算 WER, 修复了因上游 `huggingface_hub` API 变更导致的 CI 失败。

关键符号: `run_evaluation`, `run_longform_evaluation`

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。它替换了 `evaluate.load` 调用，改用 `jiwer.wer` 直接计算 WER，修复了因上游 `huggingface_hub` API 变更导致的 CI 失败。

```
# 变更前:
# from evaluate import load
# ...
# wer = load("wer")
# wer_score = 100 * wer.compute(references=references, predictions=predictions)

# 变更后:
# 使用已依赖的 jiwer 替代 evaluate，移除运行时指标下载和不兼容的 API 路径
from jiwer import wer

# 在 run_evaluation 和 run_longform_evaluation 中:
wer_score = 100 * wer(references, predictions)
```

评论区精华

无 review 评论。仅有的审核来自 `claude[bot]`（注释）和 `mgoin`（批准并致谢），没有实质性技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
 - 变更仅涉及单一测试文件中的 3 处修改（导入替换 + 两个函数中各一行简化）。
 - `jiwer` 已是测试依赖且为 Hugging Face WER 指标后端，行为等价。
 - 数值结果在 1900 个随机批次上已验证完全一致。
 - 不影响生产代码、用户 API 或其他 CI 测试。
- 影响：
 - 用户：无影响，仅改变测试内部计算方式。
 - 系统：修复持续的 CI 失败（`entrypoints-integration-speech_to_text`），恢复语音转文本正确性测试的可靠性。
 - 团队：消除了一个 CI 阻断问题，无需再处理 `evaluate` 的依赖版本冲突。通过移除 `evaluate` 的运行时下载，也减少了 CI 网络依赖。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR