PR #42274 完整报告

vllm-project/vllm

[CI] Consolidate Speech to Text tests

合并时间：2026-05-12 03:50

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42274

执行摘要

- 一句话：将 Speech-to-Text 测试独立并新增 CI 步骤
- 推荐动作：推荐作为测试基础设施清理的参考范例：测试目录与源码目录保持一致的模块化模式、CI 步骤拆分时的镜像硬件和 ignore 路径需要仔细检查、flaky test 的临时缓解方案应有后续追踪。

功能与动机

PR body 中提到 'Following #41907 Consolidate Speech to Text tests'，即跟随 #41907 的前端入口点整合工作，将测试结构对称地调整到独立目录，使测试与源码模块对齐，便于维护和 CI 调度。

实现拆解

主要步骤：

1. 在 tests/entrypoints/speech_to_text/ 下创建子目录 realtime、transcription、translation、correctness 及其 init.py、conftest.py。
2. 使用 git mv 将原有测试文件从 tests/entrypoints/openai/ 下移至新目录，并同步修改文件内的导入路径（将 from tests.entrypoints.openai.conftest 改为 from tests.entrypoints.speech_to_text.conftest）。
3. 从 tests/entrypoints/openai/chat_completion/test_enable_force_include_usage.py 中删除末尾 55 行（包含转写相关的 fixture 和测试函数），并在 tests/entrypoints/speech_to_text/transcription/ 下新建 test_enable_force_include_usage.py 文件，内容完全一致。
4. 更新 .buildkite/test-amd.yaml，为 MI300 和 MI355 硬件池分别新增 'Entrypoints Integration (Speech to Text)' CI 步骤，并从原有 Part 2 和 Part 3 命令中移除对 entrypoints/openai/speech_to_text/ 的引用和 ignore 规则。
5. 同步更新 .buildkite/test_areas/entrypoints.yaml，补充 source_file_dependencies。
6. 在 test_realtime_validation.py 的 test_multi_chunk_streaming 断言中增加一个备选 expected text，以容忍已知的 flaky transcription 输出。

关键文件：

- tests/entrypoints/speech_to_text/transcription/test_enable_force_include_usage.py（模块 测试覆盖；类别 test；类型 test-coverage；符号 transcription_server_with_force_include_usage, transcription_client_with_force_include_usage, test_transcription_with_enable_force_include_usage）：新增的独立 force-include-usage 测试文件，从 chat_completion 测试中剥离出来，正确归类到 transcription 目录。
- .buildkite/test-amd.yaml（模块 CI 配置；类别 config；类型 configuration）：CI 配置核心变更：新增专属 CI 步骤并清理旧引用。
- tests/entrypoints/speech_to_text/realtime/test_realtime_validation.py（模块 实时测试；类别 test；类型 rename-or-move）：从 openai/realtime 重命名并修改导入路径和 expected text（容忍 flaky 输出）。
- tests/entrypoints/openai/chat_completion/test_enable_force_include_usage.py（模块 测试覆盖；类别 test；类型 test-coverage；符号 transcription_server_with_force_include_usage, transcription_client_with_force_include_usage, test_transcription_with_enable_force_include_usage）：删除了末尾 55 行转写相关测试，与新增的独立测试文件对应。
- .buildkite/test_areas/entrypoints.yaml（模块 CI 区域；类别 config；类型 configuration）：同步更新 source_file_dependencies，增加 speech_to_text 目录。

关键符号：transcription_server_with_force_include_usage, transcription_client_with_force_include_usage, test_transcription_with_enable_force_include_usage

关键源码片段

tests/entrypoints/speech_to_text/transcription/test_enable_force_include_usage.py

新增的独立 force-include-usage 测试文件，从 chat_completion 测试中剥离出来，正确归类到 transcription 目录。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project

import pytest
import pytest_asyncio

from tests.utils import RemoteOpenAIServer


@pytest.fixture(scope='module')
def transcription_server_with_force_include_usage():
    args = [
        '--dtype', 'bfloat16',
        '--max-num-seqs', '4',
        '--enforce-eager',
        '--enable-force-include-usage', # 关键配置：强制在 streaming 末尾包含 usage
```

```
        '--gpu-memory-utilization', '0.2',
    ]
    with RemoteOpenAIServer('openai/whisper-large-v3-turbo', args) as remote_server:
        yield remote_server


@pytest_asyncio.fixture
async def transcription_client_with_force_include_usage(
    transcription_server_with_force_include_usage,
):
    async with (
        transcription_server_with_force_include_usage.get_async_client() as async_client
    ):
        yield async_client


@pytest.mark.asyncio
async def test_transcription_with_enable_force_include_usage(
    transcription_client_with_force_include_usage, winning_call
):
    res = await transcription_client_with_force_include_usage.audio.transcriptions.create(
        model='openai/whisper-large-v3-turbo',
        file=winning_call,
        language='en',
        temperature=0.0,
        stream=True,
        timeout=30,
    )
    # 验证每个 chunk 的 usage 字段：当 choices 为空（最终 chunk）时 usage 应为
    dict，否则不应有 usage
    async for chunk in res:
        if not len(chunk.choices):
            # final usage sent
            usage = chunk.usage
            assert isinstance(usage, dict)
            assert usage['prompt_tokens'] > 0
            assert usage['completion_tokens'] > 0
            assert usage['total_tokens'] > 0
        else:
            assert not hasattr(chunk, 'usage')
```

tests/entrypoints/speech_to_text/realtime/test_realtime_validation.py

从 openai/realtime 重命名并修改导入路径和 expected text（容忍 flaky 输出）。

```
# 关键变更：导入路径调整和 flaky 容错
from tests.entrypoints.speech_to_text.conftest import add_attention_backend

# 在 test_multi_chunk_streaming 中，原精确匹配变为容忍两种可能输出
assert full_text == (
```

```
        ' First words I spoke in the original phonograph.'
        ' A little piece of practical poetry. Mary had a little lamb,'
        ' it sleeps with quite a flow, and everywhere that Mary went,'
        ' the lamb was sure to go.'
    ) or full_text == ( # 备选结果：it squeaked with quite a flow
        ' First words I spoke in the original phonograph.'
        ' A little piece of practical poetry. Mary had a little lamb,'
        ' it squeaked with quite a flow, and everywhere that Mary went,'
        ' the lamb was sure to go.'
    )
```

评论区精华

Review 中 gemini-code-assist[bot] 指出两个 CI 配置问题：

- MI355 新增步骤的 mirror_hardwares 错误地复制了 MI300 的硬件标签（应为 amdgfx950nightly 和 amdmi355）。
- MI355 Part 3 步骤中的 --ignore=entrypoints/openai/speech_to_text/ 已是冗余路径，因为测试已移出。 作者 noooop 在 test_realtime_validation.py 中承认测试 flaky，并添加了 or full_text == ... 作为临时 workaround，表示会后续调查。
- CI 配置：MI355 步骤冗余 ignore 路径 (correctness): 冗余 ignore 应移除，与 MI300 部分的清理一致。
- CI 配置：MI355 步骤 mirror_hardwares 错误 (correctness): 修复 mirror_hardwares 为正确的 MI355 硬件标签。
- 测试 flakiness：test_multi_chunk_streaming 断言不稳定 (correctness): 接受两个可能的 transcription 输出；根本原因未解决。

风险与影响

- 风险：风险很低且可控：
  - 测试文件迁移可能导致导入路径错误或 CI 配置遗漏，但经过测试和 review 已确认正确。
  - MI355 CI 步骤的 mirror_hardwares 复制粘贴错误可能导致测试在错误硬件上运行，review 已指出，合并前应已修正。
  - test_realtime_validation.py 中的 flaky 断言通过备选文本缓解，但根本原因未解决，仍可能在特定条件下失败。
  - 影响：对用户无影响。对 CI 系统：新增独立的 Speech-to-Text 测试步骤，使测试可独立调度和快速失败定位。对团队：测试目录与源码目录结构对齐，降低维护负担。
  - 风险标记：测试组织调整，CI 配置变更，flaky test 未根本解决

关联脉络

- PR #42370 [Frontend] Consolidate Speech to Text entrypoints.: 本 PR 是 #42370 的对称操作：测试也独立到 speech_to_text 目录，实现源码与测试模块对齐。

- PR #41907 Consolidate Speech to Text tests: PR body 提及的前置 PR，触发了本次测试整合工作。具体内容未知。