

PR #43064 完整报告

vllm-project/vllm

[CI] De-flake renderers/test_hf.py::test_resolve_content_format_fallbacks[Qwen/Qwen-VL-string]

合并时间: 2026-05-22 12:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43064>

执行摘要

- 一句话: 预填充 HF 缓存修复 Qwen-VL 测试不稳
- 推荐动作: 该 PR 值得读, 因为它展示了一种处理测试中外部依赖的通用模式: 通过预填充 local cache 并结合 pytest fixture, 以最小侵入性解决 flaky 问题。设计决策 (使用 session-scoped fixture、内部 conftest) 有参考价值。

功能与动机

Qwen-VL 模型使用的 tokenizer 通过 `trust_remote_code` 加载时, 会执行 `tokenization_qwen.py` 并从 `qianwen-res.oss-cn-beijing.aliyuncs.com` 下载 `SimSun.ttf` 字体。该 CDN 在 AWS us-west-2 CI 中频繁不可达, 导致 `test_resolve_content_format_fallbacks` 测试失败 (见 Buildkite #66759)。

实现拆解

1. 新增预填充工具函数: 在 `tests/utils.py` 中定义 `prewarm_hf_cache(assets: list[tuple[str, str]]) -> None`, 内部使用 `huggingface_hub.hf_hub_download` 下载指定文件到本地 HF 缓存。函数支持 `HF_HUB_OFFLINE` 环境变量, 在离线模式跳过下载, 失败时记录警告。
2. 添加 session-scoped autouse fixture: 在 `tests/renderers/conftest.py`、`tests/tokenizers_/conftest.py` 和 `tests/models/multimodal/conftest.py` 中分别定义 `_prewarm_hf_cache` fixture, 调用 `prewarm_hf_cache` 预下载 Qwen/VL-Chat 的 `SimSun.ttf`。这些 fixture 在所有相关测试执行前自动生效。
3. 移除旧的跳过逻辑: 在 `tests/models/multimodal/processing/test_common.py` 中删除针对 Qwen/VL 和 Qwen/VL-Chat 的 `pytest.skip` 条件, 因为现在测试已能正常执行。
4. 配套导入与日志: 在 `tests/utils.py` 中添加 `huggingface_hub` 导入和 `vllm.logger` 初始化。

关键文件:

- `tests/utils.py` (模块 测试工具; 类别 test; 类型 test-coverage; 符号 `prewarm_hf_cache`): 新增了核心工具函数 `prewarm_hf_cache`, 被所有 `conftest` 调用。
- `tests/renderers/conftest.py` (模块 测试夹具; 类别 test; 类型 test-coverage; 符号 `_prewarm_hf_cache`): 新增 session-scoped fixture 调用 `prewarm_hf_cache`, 消除该目录下测试的 flaky。

- tests/tokenizers_/conftest.py (模块 测试夹具; 类别 test; 类型 test-coverage; 符号 _prewarm_hf_cache) : 与 renderers 类似, 为 tokenizers 测试提供 prewarm fixture。
- tests/models/multimodal/conftest.py (模块 测试夹具; 类别 test; 类型 test-coverage; 符号 _prewarm_hf_cache) : 为多模态模型测试添加 prewarm fixture。
- tests/models/multimodal/processing/test_common.py (模块 处理测试; 类别 test; 类型 test-coverage) : 移除了针对 Qwen-VL 的 pytest.skip, 因为预填充已解决不稳定问题。

关键符号: prewarm_hf_cache, _prewarm_hf_cache

关键源码片段

tests/renderers/conftest.py

新增 session-scoped fixture 调用 prewarm_hf_cache, 消除该目录下测试的 flaky。

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

```
import pytest
```

```
from tests.utils import prewarm_hf_cache
```

```
@pytest.fixture(scope="session", autouse=True)
def _prewarm_hf_cache():
    # tokenization_qwen.py downloads SimSun.ttf from
    # qianwen-res.oss-cn-beijing.aliyuncs.com ; both Qwen/Qwen-VL and
    # Qwen/Qwen-VL-Chat look it up from the Chat repo.
    prewarm_hf_cache([("Qwen/Qwen-VL-Chat", "SimSun.ttf")])
```

评论区精华

Review 中两个主要讨论点:

- 使用 session-scoped fixture: DarkLight1337 建议将原始的 pytest_sessionstart hook 改为 session-scoped fixture, 作者同意并修改。
- 移到内部 conftest: DarkLight1337 指出放在顶层 conftest 会影响所有测试, 建议移到内部 conftest, 作者采纳, 并指出将来新增测试需手动添加类似 fixture。
 - 使用 session-scoped fixture 替代 pytest_sessionstart hook (design): 采用 session-scoped autouse fixture, 在 conftest 中实现预填充。
 - 将 fixture 移动到内部 conftest 以减少全局影响 (design): fixture 被分散到 tests/renderers/conftest.py, tests/tokenizers_/conftest.py, tests/models/multimodal/conftest.py。

风险与影响

- 风险:

1. 外部 CDN 依赖仍然存在：预填充仅在缓存失效时下载，若 CDN 持续不可达且 HF 缓存为空，下载尝试将失败，但测试会继续（warning 而非失败）。HF_HUB_OFFLINE 环境变量可完全禁用。
2. 测试隔离性风险较低：fixture 是 session-scoped 且 autouse，但只预填充字体文件，对非 Qwen 测试无影响。
3. 维护负担：若有新的测试需类似预填充，开发者需手动添加 conftest 和调用，可能被遗忘。- 影响：直接修复 renderers/test_hf.py 中 test_resolve_content_format_fallbacks 测试的不稳定性，使其在 CI 中可靠通过。间接影响同样使用了 Qwen-VL tokenizer 的 tokenizers 测试和多模态处理测试。无用户可见的功能变化。- 风险标记：
external_cdn_dependency, test_fixture_spread

关联脉络

- PR #31932 Skip flaky Qwen-VL tests due to font download failure: 该 PR 添加了关于 Qwen-VL 测试的 pytest.skip，本 PR 移除了它并改为预填充缓存。