PR #44622 完整报告

vllm-project/vllm

[Bugfix] Update mistral tokenizer test for continue_final_message fix

合并时间：2026-06-05 16:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44622

执行摘要

- 一句话：修复 Tekken tokenizer 测试预期值
- 推荐动作：可直接合并，属于常规的测试同步更新。

功能与动机

上游 mistral-common PR#233 修复了 Tekken 分词器未转发 continue_final_message 的问题，修复后不再追加多余 EOS token，原有测试预期值过时，需要同步更新以避免测试失败。

实现拆解

1. 定位到 tests/tokenizers_/test_mistral.py 中 continue_final_message=True 的测试用例（第 798-805 行），该用例使用了 Tekken 和 V7/V15 两种 tokenizer 的 expected_output。
2. 将 Tekken tokenizer 对应的 token id 序列从 [1, 3, 22177, 4304, 2662, 4, 22177, 2] 改为 [1, 3, 22177, 4304, 2662, 4, 22177]，即移除末尾的 2（EOS token）。
3. 同时更新 decoded 字符串从 "[INST]Hello world ![/INST]Hello" 改为 "[INST]Hello world ![/INST]Hello"，移除尾部 。
4. 非 Tekken tokenizer 的预期值保持不变。

关键文件：

- tests/tokenizers_/test_mistral.py（模块 Mistral；类别 test；类型 test-coverage）：唯一修改的文件，同步更新了 continue_final_message=True 时 Tekken tokenizer 的预期 token id 列表和解码字符串，移除多余的 EOS token。

关键符号：test_apply_chat_template

关键源码片段

tests/tokenizers_/test_mistral.py

唯一修改的文件，同步更新了 continue_final_message=True 时 Tekken tokenizer 的预期 token id 列表和解码字符串，移除多余的 EOS token。

```
# tests/tokenizers_/test_mistral.py（片段）
# 变更集中在 continue_final_message=True 的测试用例
# 上游修复后，Tekken tokenizer 不再追加 EOS token（id 2），
```

```
# 因此预期 token 序列和 decode 字符串需要同步移除尾部 </s>

# 修改前（已过时）：
# expected token ids: [1, 3, 22177, 4304, 2662, 4, 22177, 2]
# decoded: "<s>[INST]Hello world ![/INST]Hello</s>"

# 修改后（正确）：
# expected token ids: [1, 3, 22177, 4304, 2662, 4, 22177]
# decoded: "<s>[INST]Hello world ![/INST]Hello"

def test_apply_chat_template(self, ...):
    actual_output = mistral_tokenizer.apply_chat_template(
        openai_request["messages"], ...
    )
    # 断言只使用对应 is_tekken 分支的预期值
    assert actual_output == expected_output[mistral_tokenizer.is_tekken]
    assert decoded_actual_output == decoded_expected_output[mistral_tokenizer.is_tekken]
```

评论区精华

无人工 review 讨论，noooop 直接 approved 并表示感谢。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试预期值，不涉及任何生产代码变更。若上游修复被回滚，测试会再次失败，但这属于预期行为，不是本 PR 引入的问题。
- 影响：影响范围仅限于 mistral tokenizer 的测试用例，确保 CI 通过；开发者无需修改使用逻辑。
- 风险标记：暂无

关联脉络

- PR #44620 [Bugfix][Rust Frontend] Fix UTF-8 char-boundary panic in incremental detokenizer: 同为 tokenizer 相关的 bugfix PR，但不在同一模块。