

PR #50241 完整报告

vllm-project/vllm

[CI][Test] Fix pooling truncation test after VLLMError hierarchy change

合并时间: 2026-07-29 20:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50241>

执行摘要

- 一句话: 修复 pooling truncation 测试异常类型不匹配
- 推荐动作: 简单但必要的测试修复, 值得快速合入以恢复 CI 稳定性。关注 PR #49665 引入的异常层次变化, 确认其他类似测试是否需要同步调整。

功能与动机

PR body 明确指出: `tests/models/language/pooling/test_truncation_control.py::test_bigger_truncation_size` 在 `main` 分支上开始失败。原因是请求校验路径现在抛出 `VLLMValidationError`, 它不再是 `ValueError` 的子类, 因此 `pytest.raises(ValueError)` 无法捕获异常, 导致测试失败。

实现拆解

1. 导入新增: 在测试文件头部添加 `from vllm.exceptions import VLLMValidationError`。
2. 异常类型替换: 将 `pytest.raises(ValueError)` 改为 `pytest.raises(VLLMValidationError)`。
3. 不涉及其他修改: 仅改动该测试文件的 2 处, 不影响任何源码逻辑。

关键文件:

- `tests/models/language/pooling/test_truncation_control.py` (模块 测试; 类别 test; 类型 test-coverage): 唯一的变更文件, 修复了因异常层次变更导致的测试失败。

关键符号: 未识别

关键源码片段

`tests/models/language/pooling/test_truncation_control.py`

唯一的变更文件, 修复了因异常层次变更导致的测试失败。

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

# 新增导入: VLLMValidationError 是 PR #49665 引入的新异常类型,
# 它继承自 VLLMClientError 而非 ValueError
from vllm.exceptions import VLLMValidationError
```

...

```
def test_bigger_truncation_size(
    vllm_runner, model_name=MODEL_NAME, input_str=input_str
):
    truncate_prompt_tokens = max_model_len + 1

    with (
        # 修复前: pytest.raises(ValueError) 无法捕获新异常
        # 修复后: 使用 pytest.raises(VLLMValidationError)
        pytest.raises(VLLMValidationError),
        vllm_runner(
            model_name, runner="pooling", max_model_len=max_model_len
        ) as vllm_model,
    ):
        llm_output = vllm_model.llm.embed(
            input_str,
            tokenization_kwargs=dict(truncate_prompt_tokens=truncate_prompt_tokens),
        )
    ...
```

评论区精华

无实质性讨论；维护者 [noooooop](#) 直接批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试文件中异常期望类型的修正，不改变任何生产代码逻辑。但需注意，若未来异常层次再次变更，该测试仍需同步更新。
- 影响：影响范围仅限于单个测试用例的通过性。修复后 ROCm CI 的 Language Models Test (Extended Pooling) 步骤恢复绿色。对其他平台无影响。
- 风险标记：仅测试修复，低风险

关联脉络

- PR #49665 [Frontend][Core] Standardize request error handling with VLLMError hierarchy: 本 PR 的修复根因：PR #49665 引入 VLLMValidationError 并迁移请求校验逻辑，导致现有测试异常捕获失效。