

PR #49655 完整报告

vllm-project/vllm

[BUGFIX] Fix log capture in KV test

合并时间: 2026-07-25 08:36

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49655>

执行摘要

- 一句话: 修复 KV 测试日志捕获问题
- 推荐动作: 快速合并。该 PR 虽小但解决了 CI 稳定性问题, 值得合入。建议关注 `caplog_vllm` fixture 的实现, 未来类似场景可直接复用。

功能与动机

PR #48776 引入的 KV 缓存变更导致一个测试用例 (`test_mla_with_incompatible_swa_uses_one_full_allocation_group`) 失败。失败的根本原因是 vLLM 的 logger 设置了 `propagate=False`, 因此 pytest 的内置 `caplog` fixture 无法捕获任何日志记录, 导致 `caplog.text` 为空, 断言 `"attention compute is unchanged" in caplog.text` 失败。

实现拆解

1. 定位问题: 通过错误信息确定 `caplog.text` 为空, 结合 vLLM 日志传播配置定位根本原因。
2. 修改 fixture: 将测试函数参数中的 `caplog` 替换为 `caplog_vllm`; `caplog_vllm` 是一个自定义 fixture, 它在使用前后临时启用日志传播到 pytest 的日志处理器。
3. 更新断言: 将断言语句中的 `caplog.text` 更新为 `caplog_vllm.text`, 确保读取的是启用了传播的日志记录。

关键文件:

- `tests/v1/core/test_kv_cache_utils.py` (模块测试; 类别 test; 类型 test-coverage; 符号 `test_mla_with_incompatible_swa_uses_one_full_allocation_group`): 唯一修改的文件, 将 `caplog` 替换为 `caplog_vllm` 修复测试失败。

关键符号: `test_mla_with_incompatible_swa_uses_one_full_allocation_group`

关键源码片段

`tests/v1/core/test_kv_cache_utils.py`

唯一修改的文件, 将 `caplog` 替换为 `caplog_vllm` 修复测试失败。

```
# tests/v1/core/test_kv_cache_utils.py
```

```
# 原代码使用 caplog fixture, 由于 vLLM 日志 propagate=False,  
# 导致 pytest 的日志捕获机制失效, caplog.text 始终为空。
```

修改后使用自定义的 caplog_vllm fixture, 它临时启用日志传播。

```
def test_mla_with_incompatible_swa_uses_one_full_allocation_group(caplog_vllm):  
    # ... 测试代码不变 ...  
    # 关键断言: 验证日志中包含了预期的警告信息  
    assert "attention compute is unchanged" in caplog_vllm.text
```

评论区精华

变更简单, 无实质性讨论。审核者 jikunshang 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低: 仅修改测试文件中的 fixture 名称, 不影响生产代码或核心逻辑。自定义 fixture 在测试使用期间启用日志传播, 不会影响其他测试。
- 影响:
 - 对用户: 无直接影响。
 - 对系统: 仅修复测试稳定性, 无功能影响。
 - 对团队: 确保 CI 中 KV 缓存相关测试稳定运行, 避免误报。
 - 风险标记: 仅测试变更

关联脉络

- PR #48776 [Bugfix][KV cache] Support sparse-MLA targets with SWA drafts: 本 PR 修复的测试用例正是由 PR #48776 引入的, 该 PR 添加了新的 KV 缓存分组逻辑, 相应的测试需要验证日志行为。