

PR #48161 完整报告

vllm-project/vllm

[CI] Increase extract hidden states TP2 timeout

合并时间: 2026-07-10 04:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48161>

执行摘要

- 一句话: 提升 `extract_hidden_states_tp2` 测试超时
- 推荐动作: 建议合并。此 PR 是轻量级的 CI 稳定性修复, 无副作用。但建议团队关注 Transformers 升级后测试时间增加的根本原因, 避免未来类似问题。

功能与动机

PR body 指出 `test_extract_hidden_states_tp2` 在 Buildkite 上因超时失败: 最后通过的主分支构建耗时 60.58s, 而后续合并 #41359 后的构建在约 62.6s 超时。此变更将 CUDA 超时从 60s 提升至 120s 以保证测试稳定。

实现拆解

在 `tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py` 中, 将 `test_extract_hidden_states_tp2` 函数的 `pytest.mark.timeout` 装饰器中条件表达式 `240 if current_platform.is_rocm() else 60` 改为 `240 if current_platform.is_rocm() else 120`。仅修改一行, 其余逻辑不变。

关键文件:

- `tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py` (模块测试; 类别 test; 类型 test-coverage): 唯一变更文件, 修改了 `test_extract_hidden_states_tp2` 的超时时间。

关键符号: `test_extract_hidden_states_tp2`

关键源码片段

`tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py`

唯一变更文件, 修改了 `test_extract_hidden_states_tp2` 的超时时间。

```
# 将超时从 60s 提升至 120s, 解决 Transformers 5.10.4 升级后的 CI 超时失败
@pytest.mark.timeout(240 if current_platform.is_rocm() else 120)
@multi_gpu_test(num_gpus=2)
@create_new_process_for_each_test()
def test_extract_hidden_states_tp2():
    """Test that hidden states extraction works with tensor_parallel_size=2."""
    # 测试体保持不变
```

评论区精华

无讨论。PR 获得 mgoin 直接批准，无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试超时参数，不影响任何业务逻辑或模型行为。但需注意，如果测试实际存在潜在性能退化，增加超时可能掩盖问题，建议后续监控测试执行时间趋势。
- 影响：影响范围仅限于 CI 中的 `test_extract_hidden_states_tp2` 测试用例，使其在 CUDA 环境下有更充裕的时间完成，避免因轻微性能波动导致 CI 误报失败。对用户无影响。
- 风险标记：CI 稳定性

关联脉络

- PR #41359 Bump Transformers version to 5.10.4: 该 PR 的合并导致了测试执行时间增加，触发了超时失败。