

# PR #49055 完整报告

vllm-project/vllm

[ROCm][CI] Ensure sliding window tests release GPU memory

合并时间: 2026-07-19 04:44

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49055>

## 执行摘要

- 一句话: 使用 `vllm_runner` 确保 GPU 内存正确释放
- 推荐动作: 建议合并。该 PR 是典型的 CI 稳定性修复, 逻辑清晰风险低, 值得学习使用上下文管理器管理引擎生命周期的模式。

## 功能与动机

ROCm CI 测试中, 滑动窗口测试用例间 GPU 内存未能及时释放, 导致下一个测试在启动时内存检查失败 (见 BuildKite 构建日志)。使用 `vllm_runner` 上下文管理器可自动触发引擎关闭和资源回收。

## 实现拆解

1. 替换引擎创建方式: 在 `tests/v1/e2e/general/test_correctness_sliding_window.py` 的 `test_sliding_window_retrieval` 函数中, 将 `LLM(...)` 的显式创建替换为 `with vllm_runner(...) as runner`: 上下文管理器。
2. 适配参数传递: `vllm_runner` 接收原传入 LLM 的模型名、`disable_hybrid_kv_cache_manager` 等参数, 并额外显式传递 `max_model_len=None` 和 `enable_chunked_prefill=None`, 确保行为一致。
3. 获取引擎引用: 通过 `runner.get_llm()` 获得 LLM 实例以供后续生成和验证使用。
4. 移除手动清理: 原有代码在函数结束前无显式释放, 上下文管理器退出时自动执行清理。

关键文件:

- `tests/v1/e2e/general/test_correctness_sliding_window.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 唯一变更文件, 替换引擎创建方式为上下文管理器以自动释放 GPU 内存。

关键符号: `test_sliding_window_retrieval`, `check_length`

## 关键源码片段

`tests/v1/e2e/general/test_correctness_sliding_window.py`

唯一变更文件, 替换引擎创建方式为上下文管理器以自动释放 GPU 内存。

```
# tests/v1/e2e/general/test_correctness_sliding_window.py
```

```
@pytest.mark.parametrize(
```

```

    "model",
    [
        "bigcode/starcoder2-3b", # sliding window only
        "google/gemma-3-1b-it", # sliding window + full attention
    ],
)
@pytest.mark.parametrize("batch_size", [5])
@pytest.mark.parametrize("seed", [1])
@pytest.mark.parametrize("disable_hybrid_kv_cache_manager", [True, False])
def test_sliding_window_retrieval(
    model, batch_size, seed, disable_hybrid_kv_cache_manager, vllm_runner # 新增 vllm_runner
    fixture
):
    # ...
    with vllm_runner(
        model,
        max_model_len=None,
        enable_chunked_prefill=None,
        disable_hybrid_kv_cache_manager=disable_hybrid_kv_cache_manager,
        enforce_eager=enforce_eager,
    ) as runner: # 上下文管理器自动管理 LLM 生命周期
        llm = runner.get_llm()
        sampling_params = SamplingParams(temperature=0.0, max_tokens=100)
        prompts, answer, indices = prep_prompts(
            batch_size, ln_range=test_config.ln_range
        )
        # ... 生成与断言逻辑不变 ...
    # 退出 with 块后 vllm_runner 自动关闭引擎并释放 GPU 内存

```

## 评论区精华

无讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。变更仅限于测试文件，仅修改了引擎创建方式，业务逻辑和断言未变。  
vllm\_runner 已在其他测试中广泛使用，行为稳定。
- 影响：影响范围仅限 ROCm CI 环境下的滑动窗口测试，提升测试稳定性和可靠性，对其他测试无影响。
- 风险标记：测试稳定性

## 关联脉络

- 暂无明显关联 PR