PR #41102 完整报告

vllm-project/vllm

[CI] Stabilize cpu offload compressed tensors test

合并时间：2026-05-04 13:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41102

执行摘要

- 一句话：修复 CPU offload 压缩张量测试的不稳定性
- 推荐动作：该 PR 属于小型测试稳定性修复，适合快速合并。可关注后续是否需要对其他测试也应用类似模式。

功能与动机

Avoid flaky stochastic sampling in the compressed-tensors MoE CPU offload test. 通过禁用 temperature=1 的 seeded sampling 部分，消除随机性带来的测试不稳定。

实现拆解

1. 修改 tests/utils.py 中的 _test_completion 函数：新增 include_seeded_sampling: bool = True 参数，当参数为 False 时跳过 seeded sampling 测试块。
2. 修改 tests/utils.py 中的 compare_two_settings 和 compare_all_settings 函数：同样新增 include_seeded_sampling 参数并透传到 _test_completion，使得上层调用可以控制是否执行 seed 采样检测。
3. 修改 tests/quantization/test_cpu_offload.py 中的 test_cpu_offload_compressed_tensors 测试：在调用 compare_two_settings 时传入 include_seeded_sampling=False，显式禁用不稳定 seed 采样部分，其余确定性检验（贪婪、token IDs、list、流式）保持不变。

关键文件：

- tests/utils.py（模块 测试工具；类别 test；类型 test-coverage）：核心修改文件：新增 include_seeded_sampling 参数，使 _test_completion、compare_two_settings、compare_all_settings 可选择性跳过 seed 采样测试，是测试稳定性的关键控制点。
- tests/quantization/test_cpu_offload.py（模块 CPU 卸载；类别 test；类型 test-coverage）：测试用例修改：在 test_cpu_offload_compressed_tensors 中调用 compare_two_settings 时传入 include_seeded_sampling=False，禁用随机 seed 采样检测，提升测试稳定性。

关键符号：_test_completion, compare_two_settings, compare_all_settings

关键源码片段

tests/utils.py

核心修改文件：新增 include_seeded_sampling 参数，使 _test_completion、compare_two_settings、compare_all_settings 可选择性跳过 seed 采样测试，是测试稳定性的关键控制点。

```
def _test_completion(
    client: openai.OpenAI,
    model: str,
    prompt: str,
    token_ids: list[int],
    include_seeded_sampling: bool = True, # 新增：控制是否执行 seed 采样测试
):
    results = []
    # ... 其他确定性测试（greedy, token_ids, list, streaming）始终执行 ...

    if include_seeded_sampling:
        # 原来无条件的 seed 采样现在被包裹在条件判断中
        completion = client.completions.create(
            model=model, prompt=prompt, max_tokens=5, seed=33, temperature=1.0
        )
        results.append({"test": "seeded_sampling", ...})

        # 多 prompt 的 seed 采样
        completion = client.completions.create(
            model=model, prompt=[prompt, prompt], max_tokens=5, seed=33, temperature=1.0
        )
        results.append({"test": "seeded_sampling", ...})

    return results
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅影响测试辅助函数与单个测试用例，不涉及生产代码。修改后其他测试若默认使用 include_seeded_sampling=True，行为不变；没有回归风险。
- 影响：影响范围局限于 test_cpu_offload_compressed_tensors 测试：移除了 seed 采样检测，减少因随机性导致的 CI 不稳定。其他测试不受影响。对用户和系统无影响。
- 风险标记：无生产代码变更，测试重要性低

关联脉络

- 暂无明显关联 PR