PR #49351 完整报告

vllm-project/vllm

[CI][Bugfix] Reduce max_model_len in OOT embedding test to fix KV-cache OOM on small GPUs

合并时间：2026-07-22 03:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49351

执行摘要

- 一句话：降低 embedding 测试的 max_model_len 修复 OOM
- 推荐动作：可以直接合并。这是一个简单且安全的测试修复，通过调整测试配置避免 CI 硬件限制导致的失败。

功能与动机

关联 Issue #49333 报告了 CI 测试 test_oot_registration_embedding 在 L4 GPU 上因 KV cache 内存不足而失败：模型权重和 CUDA graph 内存预留 (0.50 GiB) 后只剩下 0.64 GiB，而 max_model_len=2048 需要 0.66 GiB。PR body 确认测试只嵌入两个短 prompt，实际不需要长序列，因此通过降低 max_model_len 解决。

实现拆解

1. 在 tests/plugins_tests/test_oot_registration_offline.py 的 test_oot_registration_embedding 函数中，将 LLM 初始化参数 max_model_len 从 2048 改为 512。
2. 这个改动使 KV cache 需求从 0.66 GiB 降至约 0.17 GiB，确保在 L4 GPU 的剩余 0.64 GiB KV cache 内存内可以启动。
3. 由于测试只嵌入两个约 7-token 的 prompt，max_model_len=512 完全满足测试需求，不会改变测试语义或输出。

关键文件：

- tests/plugins_tests/test_oot_registration_offline.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，将 max_model_len 从 2048 降低到 512，修复 KV cache OOM。

关键符号：test_oot_registration_embedding

关键源码片段

tests/plugins_tests/test_oot_registration_offline.py

唯一变更文件，将 max_model_len 从 2048 降低到 512，修复 KV cache OOM。

```
# tests/plugins_tests/test_oot_registration_offline.py
@create_new_process_for_each_test()
```

```
def test_oot_registration_embedding(
    monkeypatch: pytest.MonkeyPatch,
    dummy_gemma2_embedding_path: str,
):
    with monkeypatch.context() as m:
        m.setenv("VLLM_PLUGINS", "register_dummy_model")
        prompts = ["Hello, my name is", "The text does not matter"]
        llm = LLM(
            model=dummy_gemma2_embedding_path,
            load_format="dummy",
            # 降低 max_model_len 以减少 KV cache 需求，避免小 GPU 上 OOM
            # 原值 2048 需要 0.66 GiB KV cache，但 L4 GPU 只剩 0.64 GiB
            max_model_len=512, # 只需 0.17 GiB，远低于可用内存
        )
        outputs = llm.embed(prompts)

        for output in outputs:
            assert all(v == 0 for v in output.outputs.embedding)
```

评论区精华

无人工 review 评论；yewentao256 直接批准，claude[bot] 自动评论因来自 fork 而跳过检查。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改测试参数，测试逻辑和断言不变。若未来某个 embedding 模型需要更长的序列进行测试，该参数可能需要重新调整，但当前测试场景不涉及长序列。
- 影响：仅影响 CI 中的 test_oot_registration_embedding 测试，使其能在 L4 等小显存 GPU 上通过。不改变任何生产代码或 API 行为。
- 风险标记：测试配置变更，硬件相关失败

关联脉络

- PR #38284 Enable CUDA graph memory profiling by default: 该 PR 默认启用了 CUDA graph memory profiling，导致 KV cache 可用内存减少 0.50 GiB，间接引发本 PR 修复的 OOM 问题。
- PR #49333 [CI Failure]: Plugin Tests (2 GPUs): 关联 Issue，详细描述了失败的根因和内存计算链路。