PR #41147 完整报告

vllm-project/vllm

[CI] De-flake test_chat_completion_n_parameter_non_streaming

合并时间：2026-04-29 11:24

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41147

执行摘要

- 一句话：修复 chat 测试的 n 参数 flaky 问题
- 推荐动作：值得合并以提升 CI 稳定性。

功能与动机

PR body 指出测试断言 len(set(contents)) > 1 在低熵 prompt 和短输出下可能全部相同（见 Buildkite #63256），需要使 n 个样本的多样性变为确定性。

实现拆解

仅修改 tests/entrypoints/openai/chat_completion/test_chat.py 中 test_chat_completion_n_parameter_non_streaming 函数的三处参数：

1. 增加 seed=42 使采样确定性，vLLM 自动为 n=3 衍生子种子 42/43/44，保证不同。
2. temperature 从 0.7 提高到 1.0 使采样分布更均匀，减少偶然雷同。
3. max_completion_tokens 从 20 扩大到 50，提高文本长度丰富度。
4. 移除之前的注释 # Verify all responses are different (highly likely with temperature > 0)，但未补充新注释。

关键文件：

- tests/entrypoints/openai/chat_completion/test_chat.py（模块 测试；类别 test；类型 test-coverage；符号 test_chat_completion_n_parameter_non_streaming）：唯一修改的文件，修复 flaky 测试的三个参数和移除注释。

关键符号：test_chat_completion_n_parameter_non_streaming

关键源码片段

tests/entrypoints/openai/chat_completion/test_chat.py

唯一修改的文件，修复 flaky 测试的三个参数和移除注释。

```
    # Test with n=3
    chat_completion = await client.chat.completions.create(
        model=model_name,
        messages=messages,
        max_completion_tokens=50, # Changed from 20 to 50
```

```
        temperature=1.0, # Changed from 0.7 to 1.0
        n=3,
        seed=42, # Added seed for deterministic diversity
        stream=False,
    )

    assert len(chat_completion.choices) == 3

    # Verify each choice has content and correct index
    for i, choice in enumerate(chat_completion.choices):
        assert choice.index == i
        assert choice.message.content is not None
        assert len(choice.message.content) > 0

    # Removed previous comment "Verify all responses are different (highly likely with
    temperature > 0)"
    contents = [choice.message.content for choice in chat_completion.choices]
    assert len(set(contents)) > 1, "Expected different responses with n=3"
```

评论区精华

gemini-code-assist[bot] 建议恢复或更新注释以反映确定性种子逻辑，但实际未采纳。DarkLight1337 直接批准。

- 恢复断言注释 (documentation): 注释被移除，未重新添加。

风险与影响

- 风险：仅修改测试参数，无生产代码风险。但若模型内部种子衍生产生改变，此测试可能仍然失败——不过 vLLM 的种子逻辑较稳定。
- 影响：直接影响 CI 可靠性，消除一个已知 flaky 测试。对其他模块无影响。
- 风险标记：仅测试变更

关联脉络

- PR #41121 [CI] Return HTTP 400 for unsupported chat content part type: PR author 在 Issue 评论中指出此 flaky 可能由另一已知问题 PR#41121 引起。