PR #42392 完整报告

vllm-project/vllm

[CI] De-flake Language Models Test (Extended Generation)
test_models(False-False-5-32-bigcode/starcoder2-3b)

合并时间：2026-05-12 18:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42392

执行摘要

- 一句话：修复 starcoder2-3b 测试 flaky 问题
- 推荐动作：值得精读。这是一个典型的因浮点精度差异导致的测试 flaky 修复方案：通过调整输入 prompt 使模型输出更稳定，而非放宽测试断言。体现了对问题根源的跟踪和分析。

功能与动机

长期存在的 CI 失败，issue #37304 和 #42336 追踪。starcoder2-3b 是代码模型，自然语言 prompt 在约 8 个 token 匹配后进入 Jupyter markdown 格式，导致 logit 分布接近均匀，HF 与 vLLM 的 bf16 计算差异引起 top-K 排序翻转，测试断言失败。

实现拆解

1. 在 tests/models/language/generation/test_common.py 的 test_models 函数中，在触发 HF runner 之前，特判 model == "bigcode/starcoder2-3b"。
2. 将 example_prompts[1]（原 Test1 的自然语言 prompt）替换为一个 Python 函数定义的代码 prompt `def add(a, b):\n return a + b

def sub(a, b):\n return a - `。

1. 替换使用 list(example_prompts) 创建副本，避免修改夹具共享的资源。其他模型保持原 prompt 不变。

关键文件：

- tests/models/language/generation/test_common.py（模块 测试；类别 test；类型 test-coverage）：核心修复文件：为 starcoder2-3b 模型特例化 Test1 prompt 为代码 prompt，消除 logprobs 排序不一致导致的 flaky。

关键符号：未识别

关键源码片段

tests/models/language/generation/test_common.py

核心修复文件：为 starcoder2-3b 模型特例化 Test1 prompt 为代码 prompt，消除 logprobs 排序不一致导致的 flaky。

```
# tests/models/language/generation/test_common.py
# 在 test_models 函数中，添加如下片段（位置在 AITER 相关条件之后，HF runner 之前）：
if model == "bigcode/starcoder2-3b":
    # 将 example prompts 转为列表以修改（原为 tuple，不可变）
    example_prompts = list(example_prompts)
    # 用代码 prompt 替换 Test1（索引 1），使模型处于训练分布内
    # 原因：starcoder2-3b 是代码模型，NL prompt 导致 logit 近似均匀，
    # bf16 舍入误差会改变 top-K 排序，引发 vLLM 与 HF 结果不一致。
    example_prompts[1] = (
        "def add(a, b):\n    return a + b

def sub(a, b):\n return a - "
    )
```

评论区精华

无 review 讨论。两个 bot 自动评论，人类 reviewer DarkLight1337 和 ZJY0516 均 Approve。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅对特定模型（starcoder2-3b）更改了测试 prompt，不影响其他测试或生产代码。修改在测试函数内局部生效，不会干扰其他模型。
- 影响：直接影响长期 flaky 的 CI 测试，预计将稳定 L4 上的 starcoder2-3b 测试。不影响用户、系统功能或性能。
- 风险标记：暂无

关联脉络

- PR #34644 [release 2.11] Update to torch 2.11: PyTorch 2.11 升级引入 bf16 行为变化，触发了该测试失败。
- PR #37304 [Bug]: Language Models Test (Extended Generation) test_models[False-False-5-32-bigcode/starcoder2-3b] test issue: 跟踪该测试失败的原始 issue。
- PR #42336 [CI Failure]: Language Models Test (Extended Generation): 最近的 CI 失败报告，用于精确定位回归窗口。