# PR #49511 完整报告

vllm-project/vllm

[CI] Disable reasoning in Responses smoke test

合并时间：2026-07-24 03:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49511

## 执行摘要

- 一句话：禁用 Responses 冒烟测试中的推理输出
- 推荐动作：简单明确的 CI 稳定性修复，无需深入阅读。

## 功能与动机

Responses API 冒烟测试中 Qwen3 可能将输出预算用于推理轨迹，导致响应标记为 incomplete。该测试只检查基本响应完成，因此显式禁用推理以确保稳定性。

## 实现拆解

在 tests/entrypoints/openai/responses/test_simple.py 的 test_basic 函数中，向 client.responses.create 调用添加了 reasoning={"effort": "none"} 参数，一行改动。

关键文件：

- tests/entrypoints/openai/responses/test_simple.py（模块 测试脚本；类别 test；类型 test-coverage）：单行修改，在 test_basic 请求中添加 reasoning={'effort': 'none'} 以禁用推理输出

关键符号：test_basic

## 关键源码片段

### tests/entrypoints/openai/responses/test_simple.py

单行修改，在 test_basic 请求中添加 reasoning={'effort': 'none'} 以禁用推理输出

```
# tests/entrypoints/openai/responses/test_simple.py

@pytest.mark.asyncio
@pytest.mark.parametrize("model_name", [MODEL_NAME])
async def test_basic(client: OpenAI, model_name: str):
    response = await client.responses.create(
        model=model_name,
        input="What is 123 * 456?",
        reasoning={"effort": "none"}, # 新增：显式禁用推理，避免输出预算被推理轨迹消耗
    )
    assert response is not None
```

```
print("response: ", response)
assert response.status == "completed"
assert response.incomplete_details is None
```

## 评论区精华

无需高阶设计讨论。CI 机器人自动评论（fork 限制），审核者 chaunceyjiang 直接批准。

- 暂无高价值评论线程

## 风险与影响

- 风险：无风险。仅修改测试用例参数，不影响生产逻辑。
- 影响：影响范围极小，仅限于 test_basic 测试函数。解决了 CI 中 Entrypoints Integration (Responses API) 测试组的 flaky 问题。
- 风险标记：测试稳定性

## 关联脉络

- 暂无明显关联 PR