PR #44618 完整报告

vllm-project/vllm

[Bugfix] Fix test_invocations flaky failure with newer openai SDK

合并时间：2026-06-05 15:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44618

执行摘要

- 一句话：修复 test_invocations 因新版 openai SDK 的不稳定失败
- 推荐动作：值得快速合并的 bugfix。虽改动简单，但解决了因外部依赖升级引起的测试不稳定问题，提升了 CI 可靠性。

功能与动机

修复因 openai SDK 版本升级（>=2.32）导致的 test_invocations 偶发失败：SDK 的 model_dump() 会注入客户端侧字段（如 moderation），这些字段不存在于 /invocations 原始响应中，导致键比较断言 chat_output.keys() == invocation_output.keys() 失败。

实现拆解

1. 定位问题根源：test_invocations 函数原先使用 client.chat.completions.create()（异步 SDK）获取 chat completion，再调用 model_dump() 转换为 dict；而 invocations 端则使用 requests.post() 直接调用原始 HTTP 接口。两个路径经不同的序列化逻辑，导致响应 dict 的键集合不一致。
2. 统一请求方式：将 chat completion 也改为使用 requests.post() 直接发送 HTTP 请求到 server.url_for("v1/chat/completions")，并将响应解析为 chat_response.json()。这样两个响应都来自服务器原始 JSON，键结构完全一致。
3. 调整变量名与断言：将 chat_completion 改为 chat_response，chat_completion.model_dump() 改为 chat_response.json()，保持后续键比较和 choices 比较逻辑不变。
4. 添加注释说明：在变更处添加注释，解释为何使用原始 HTTP 而非 SDK，帮助后续维护者理解。
5. 改动文件：仅修改 tests/entrypoints/openai/chat_completion/test_chat.py 中的 test_invocations 函数，+8/-2。

关键文件：

- tests/entrypoints/openai/chat_completion/test_chat.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，修改了 test_invocations 函数，将 chat completion 请求从 SDK 调用改为原始 HTTP 请求，消除 SDK 版本差异导致的测试不稳定。

关键符号：test_invocations

关键源码片段

tests/entrypoints/openai/chat_completion/test_chat.py

唯一变更文件，修改了 test_invocations 函数，将 chat completion 请求从 SDK 调用改为原始 HTTP 请求，消除 SDK 版本差异导致的测试不稳定。

```
# 在 test_invocations 函数中，将原先使用 openai SDK 的异步调用
# 改为使用 requests.post() 直接发送 HTTP 请求，使得两个端点
# (v1/chat/completions 和 invocations) 的响应都来自服务器原始 JSON，
# 避免 SDK model_dump() 注入额外字段 ( 如 moderation) 导致键不匹配。

# 原代码 (base):
# chat_completion = await client.chat.completions.create(**request_args)
# ...
# chat_output = chat_completion.model_dump()

# 新代码 (head):
# 使用原始 HTTP 请求，与下方 invocations 保持一致
chat_response = requests.post(
    server.url_for("v1/chat/completions"), json=request_args
)
chat_response.raise_for_status()

invocation_response = requests.post(
    server.url_for("invocations"), json=request_args
)
invocation_response.raise_for_status()

chat_output = chat_response.json()
invocation_output = invocation_response.json()

# 后续断言不变
assert chat_output.keys() == invocation_output.keys()
assert chat_output["choices"] == invocation_output["choices"]
```

评论区精华

无 review 评论讨论；PR 已被合入者 noooop 直接批准，并确认问题可在 openai SDK 从 2.40.0 升级至 2.41.0 时本地复现。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。改动仅限测试文件，且逻辑等价——两端现在都使用原始 HTTP 请求，不改变任何生产代码。但需确保 server.url_for("v1/chat/completions") 路径格式与服务器端路由匹配；若路径拼写错误，测试会失败，但不会影响生产。

- 影响：影响范围限定于 test_invocations 单个测试用例。修复后该测试不再受 openai SDK 版本影响，消除了 CI 中的不稳定性。对用户无影响，对开发者来说测试更可靠。
- 风险标记：测试稳定性，外部依赖兼容

关联脉络

- 暂无明显关联 PR