

PR #47250 完整报告

vllm-project/vllm

[Test] Run SageMaker handler-override tests in-process via TestClient

合并时间: 2026-07-01 17:14

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47250>

执行摘要

- 一句话: SageMaker handler 覆盖测试改用 TestClient, 移除模型服务器依赖
- 推荐动作: 值得关注的是其 in-process 测试模式: 针对仅测试非推理逻辑的场景, 用 TestClient 替代全服务器启动可显著提高 CI 可靠性和速度。但需确认 attach_router 和 sagemaker_standards_bootstrap 的调用方式与生产一致, 避免测试环境与生产环境的行为偏差。

功能与动机

原有测试启动完整模型服务器以测试 handler 覆盖逻辑, 但覆盖逻辑完全替换端点, 不需要模型。启动服务器不仅慢, 而且对 FastAPI 版本敏感: 解析到 fastapi ≥ 0.137 时路由树变化会导致测试失败。因此改为进程内测试以提升速度和确定性。

实现拆解

实现步骤如下:

1. 新建 `_build_sagemaker_test_client` 辅助函数: 在测试文件头部定义, 内部调用 `vllm.entrypoints.serve.sagemaker.api_router` 的 `attach_router` 和 `sagemaker_standards_bootstrap`, 构建一个真实的 FastAPI 应用并包装成 TestClient。 `attach_router` 传入空任务列表, 因为覆盖测试不依赖任何任务处理器。
2. 重写六个覆盖测试: 每个测试原用 `async def` 并依赖 `RemoteOpenAIServer`, 现改为同步 `def`, 使用 `monkeypatch fixture` 设置环境变量, 使用 `tmp_path fixture` 创建临时脚本文件。测试通过 `_build_sagemaker_test_client()` 获取客户端, 对 `/ping` 和 `/invocations` 发送请求验证响应。
3. 保留 `test_framework_default_handlers`: 该测试真正执行默认推理路径, 因此仍使用 `RemoteOpenAIServer` 启动模型服务器。未修改此测试。
4. 清理和简化: 移除 `import tempfile` 和 `pytest.mark.asyncio` 装饰, 删除 `setup_method` 中不必要的缓存清除 (因为 TestClient 在每次调用时重建应用实例, 状态不共享)。

关键文件:

- `tests/entrypoints/serve/sagemaker/test_sagemaker_handler_overrides.py` (模块 覆盖测试; 类别 test; 类型 test-coverage; 符号 `_build_sagemaker_test_client`, `test_customer_script_functions_auto_loaded`, `test_customer_decorator_usage`, `test_handler_priority_order`): 唯一变更文件, 重写了所有 handler-override 测试的实现

方式，从启动远程服务器改为进程内 TestClient，同时保留一个需要真实推理的测试不变。

关键符号：_build_sagemaker_test_client, test_customer_script_functions_auto_loaded, test_customer_decorator_usage, test_handler_priority_order, test_environment_variable_script_loading, test_handler_env_var_override, test_env_var_priority_over_decorator_and_script

关键源码片段

[tests/entrypoints/serve/sagemaker/test_sagemaker_handler_overrides.py](#)

唯一变更文件，重写了所有 handler-override 测试的实现方式，从启动远程服务器改为进程内 TestClient，同时保留一个需要真实推理的测试不变。

```
def _build_sagemaker_test_client() -> TestClient:
    """Build a TestClient over the real SageMaker router and bootstrap path.

    ``attach_router`` is called with empty supported tasks because the override
    tests replace the endpoints with customer handlers, so no framework
    invocation handler (and therefore no engine) is exercised.
    """
    from vllm.entrypoints.serve.sagemaker.api_router import (
        attach_router,
        sagemaker_standards_bootstrap,
    )

    app = FastAPI()
    attach_router(app, ())
    return TestClient(sagemaker_standards_bootstrap(app))


class TestHandlerOverrideIntegration:

    def test_customer_script_functions_auto_loaded(
        self, monkeypatch: pytest.MonkeyPatch, tmp_path: pathlib.Path
    ):
        from model_hosting_container_standards.sagemaker.config import SageMakerEnvVars

        # 创建客户自定义脚本
        script_path = tmp_path / "model.py"
        script_path.write_text(
            """
from fastapi import Request

async def custom_sagemaker_ping_handler():
    return {"status": "healthy", "source": "customer_override", "message": "Custom ping from customer script"}

async def custom_sagemaker_invocation_handler(request: Request):
    return {"predictions": ["Custom response from customer script"], "source": "customer_
```

```

        override"}
    ...

    )

    # 模拟 SageMaker 环境变量
    monkeypatch.setenv(SageMakerEnvVars.SAGEMAKER_MODEL_PATH, str(tmp_path))
    monkeypatch.setenv(SageMakerEnvVars.CUSTOM_SCRIPT_FILENAME, script_path.name)

    # 使用真正的引导路径，无需启动服务器
    client = _build_sagemaker_test_client()
    ping_resp = client.get("/ping")
    assert ping_resp.status_code == 200
    assert ping_resp.json()["source"] == "customer_override"

    invoke_resp = client.post("/invocations", json={"inputs": "test"})
    assert invoke_resp.status_code == 200
    assert invoke_resp.json()["source"] == "customer_override"

```

评论区精华

本 PR 未产生实质性 review 讨论。[DarkLight1337](#) 直接批准，[claude\[bot\]](#) 自动评论但未开启 fork 的自动 review。

- 暂无高价值评论线程

风险与影响

- 风险：仅修改测试文件，未改动任何生产代码或项目配置，引入回归风险极低。六个测试全部通过（已附上测试结果）。唯一潜在风险是若未来 `attach_router` 或 `sagemaker_standards_bootstrap` 接口变更，这些测试可能过时，但因为是直接调用生产函数，反而能较早暴露这类不兼容。
- 影响：影响范围仅限于 SageMaker handler 覆盖集成测试的六个用例。改动后测试速度大幅提升（原需启动模型服务器的 1.77s vs 原先可能分钟级），且消除对 FastAPI 版本的脆弱依赖。对其他模块、用户或系统无影响。
- 风险标记：暂无

关联脉络

- PR #44194 [Test] Convert SageMaker handler-override test to TestClient (abandoned): 之前的尝试，仅转换了一个测试，被放弃。本 PR 转换了全部六个服务器无关的覆盖测试。