

PR #46482 完整报告

vllm-project/vllm

[ROCm][P/D] MoRIIO toy proxy: support JSON Content-Type for OpenAI clients.

合并时间: 2026-07-02 03:17

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46482>

执行摘要

- 一句话: 修复 MoRIIO 代理 Content-Type 缺失
- 推荐动作: 该 PR 为一次简单但必要的 bugfix, 建议快速合并。虽然改动很小, 但解决了客户端兼容性的根本问题, 值得阅读以了解 Quart 代理的常见陷阱。

功能与动机

Quart 流式响应默认 Content-Type 为 `text/html; charset=utf-8`, 导致 OpenAI 兼容 JSON 客户端 (如 `lm_eval`) 调用 `response.json()` 时因 `mimetype` 不匹配抛出 `ContentTypeError`, 所有请求返回空结果。PR body 指出: 'The result is (no outputs) for every request and no accuracy score, even though the prefill/decode disaggregation path itself works correctly.' 该代理是唯一缺失此头的 Quart 代理, 需要对齐以支持客户端正常工作。

实现拆解

1. 定位问题: 在 `moriiio_toy_proxy_server.py` 文件的 `handle_request` 函数中, 创建响应对象后未设置 Content-Type 头, Quart 默认使用 `text/html`。
2. 添加 Content-Type 设置: 在 `response = await make_response(stream_generator)` 之后插入一行: `response.headers["Content-Type"] = decode_response.headers.get("Content-Type", "application/json")`。该行从后端解码实例的响应头中获取 Content-Type, 若后端未设置则默认使用 `application/json`。
3. 影响范围: 仅修改一个文件, 新增 3 行代码, 无额外测试, 但对 OpenAI 客户端兼容性至关重要。

关键文件:

- `examples/disaggregated/disaggregated_serving/moriiio_toy_proxy_server.py` (模块 代理服务器; 类别 `source`; 类型 `core-logic`): 核心修复文件, 在 `handle_request` 函数中为流式响应添加 Content-Type 头。

关键符号: `handle_request`

关键源码片段

`examples/disaggregated/disaggregated_serving/moriiio_toy_proxy_server.py`

核心修复文件, 在 `handle_request` 函数中为流式响应添加 Content-Type 头。

```
# ... 之前的代码
session, decode_response = await decode_request_task
stream_generator = stream_decode_response(session, decode_response, request_id)
response = await make_response(stream_generator)

# 关键修复：设置 Content-Type 响应头
# 从后端解码实例的响应头中获取 Content-Type，若后端未设置则使用 "application/json"
# 这确保了 OpenAI 兼容 JSON 客户端（如 lm_eval）能够正确解析响应
response.headers["Content-Type"] = decode_response.headers.get(
    "Content-Type", "application/json"
)
return response
# ... 异常处理
```

评论区精华

PR 获得两位 reviewer 批准，无 review 评论。讨论主要集中在 PR 描述中问题复现和修复的必要性。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅改动一行，添加响应头赋值，无逻辑分支变化。若后端返回非 JSON Content-Type，该头会被正确传递；默认 application/json 与 OpenAI 协议一致。对非 JSON 客户端的潜在影响很小。
- 影响：直接影响使用 MoRIIO 代理的 OpenAI 兼容客户端（如 lm_eval），使其能够正确解析 JSON 响应。对其他用户无影响，因为该修复只修正了 HTTP 头的语义，不改变响应体内容。影响范围局限于去中心化服务示例场景。
- 风险标记：暂无

关联脉络

- PR #42748 [Bugfix] Expose usage field in GenerateResponse for disaggregated serving: 同样涉及去中心化服务代理的响应兼容性问题，属于同一功能领域。