PR #38519 完整报告

vllm-project/vllm

Fix Responses JSON schema alias serialization

合并时间：2026-04-09 10:50

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38519

执行摘要

本 PR 修复了 vLLM Responses API 中 JSON Schema 序列化的 bug，通过添加 by_alias=True 参数确保 Pydantic 模型字段别名正确使用，避免了内部字段 schema_ 泄漏，恢复了非 Harmony tool-calling 功能。变更影响前端 Responses 模块，涉及 API 路由、协议和测试更新，是一个有意义的改进，值得关注序列化设计决策。

功能与动机

为什么做：Responses API 在使用 JSON Schema 时，非流式响应和流式响应会错误地输出内部字段 schema_ 而非公开字段 schema，导致非 Harmony tool-calling（当 tool_choice="required" 时）中断。Issue #38245 报告了此 bug，PR #38262 尝试修复但未完全解决流式响应问题。

实现拆解

按模块拆解改动：

- API 路由模块(vllm/entrypoints/openai/responses/api_router.py)：在 _convert_stream_to_sse_events、create_responses、retrieve_responses 和 cancel_responses 函数中，为 model_dump_json 和 model_dump 添加 by_alias=True，确保事件流和 JSON 响应序列化使用别名。
- 协议模块(vllm/entrypoints/openai/responses/protocol.py)：修改 serialize_message 函数，用 model_dump(mode="json", by_alias=True) 替代 model_dump_json(by_alias=True)，返回 Python 字典而非 JSON 字符串，修复消息序列化回归。
- 服务模块(vllm/entrypoints/openai/responses/serving.py)：在生成 ResponseCreatedEvent 时添加 by_alias=True，确保初始响应正确序列化。
- 测试模块(tests/entrypoints/openai/responses/test_serving_responses.py)：新增 test_response_created_event_uses_public_json_schema_alias 测试，验证别名序列化；更新 Harmony 相关测试使用 Message.from_dict 处理消息格式。

评论区精华

提炼 review 讨论：

- 在 Issue 评论中，noobHappylife 和 chaunceyjiang 发现了 Harmony 消息序列化不一致问题：非流式测试假设嵌套 author 格式，而流式测试使用 Message.from_dict 期望扁平格式。

chaunceyjiang 确认：> " 使用嵌套 author 结构会导致数据丢失。" 最终结论是更新测试以使用 Message.from_dict，确保消息格式正确。

- Review 中，gemini-code-assist[bot] 总结变更：> " 更新序列化以使用字段别名，确保如 schema 字段正确命名。" chaunceyjiang 批准 PR。

风险与影响

具体风险：

1. 序列化逻辑变更：API 响应格式可能因其他未处理别名字段而受影响，但变更局限于 Responses 模块，风险较低。
2. 回归风险：serialize_message 函数修改返回类型，可能影响依赖组件，但添加了单元测试 test_serialize_message_pydantic_model_returns_dict 覆盖。
3. 兼容性影响：修复后，API 响应将严格遵循 OpenAI 规范，提升工具调用兼容性，对用户透明。

影响范围：

- 用户：修复 tool-calling 中断，提升使用 JSON Schema 和 tool_choice="required" 场景的体验。
- 系统：确保 Responses API 输出一致性，减少字段泄漏，增强可靠性。
- 团队：测试更新强调了消息格式处理，为类似序列化问题提供参考。

关联脉络

与历史 PR 的关系：

- PR #38262（"[frontend] dump openai responses type by alias"）直接关联，它尝试修复同一 Issue 但未解决流式响应问题，本 PR 是后续完善。
- 结合近期 PR 分析，如 PR #39114（"[Bugfix] Fix Gemma4 streaming tool call corruption"）也涉及 tool-calling 修复，显示团队在前端和工具调用模块持续优化兼容性和稳定性。本 PR 是这一趋势的一部分，专注于 API 响应序列化细节，以确保与 OpenAI 标准对齐。