PR #46843 完整报告

vllm-project/vllm

[Bugfix][Parser] Pass token IDs to parser.parse() in Responses API and batch serving

合并时间：2026-06-27 03:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46843

执行摘要

- 一句话：补齐 parser.parse() 的 token_ids 参数传递
- 推荐动作：变更明确且安全，已获批准。关注后续解析器迁移 PR 以了解 token_ids 将如何被利用。

功能与动机

PR body 指出："Chat completion serving already passes model_output_token_ids to parser.parse(), but Responses API and batch serving did not. Wire those call sites to pass token IDs consistently, so that all non-streaming calls to parser.parse() provide token ids and we get consistent parsing logic across all based on those token ids." 这是为后续将解析器从 DelegatingParser 迁移到统一解析器所做的准备性工作。

实现拆解

1. 批量服务入口：在 vllm/entrypoints/openai/chat_completion/batch_serving.py 的 parser.parse() 调用中增加 model_output_token_ids=output.token_ids 参数。
2. Responses API 上下文：在 vllm/entrypoints/openai/responses/context.py 的 append_output() 方法中增加 model_output_token_ids=completion.token_ids 参数。
3. Responses API 输出构造：在 vllm/entrypoints/openai/responses/serving.py 的 _make_response_output_items() 方法中增加 model_output_token_ids=final_output.token_ids 参数。 每个变更点均只增加一行，不改变已有调用逻辑或返回值。

关键文件：

- vllm/entrypoints/openai/chat_completion/batch_serving.py（模块 入口层；类别 source；类型 core-logic）：批量服务中 parser.parse() 调用点，缺少 token_ids 参数。
- vllm/entrypoints/openai/responses/context.py（模块 入口层；类别 source；类型 core-logic）：Responses API 上下文中的解析调用点，缺少 token_ids 参数。
- vllm/entrypoints/openai/responses/serving.py（模块 入口层；类别 source；类型 core-logic）：Responses API 输出构造中的解析调用点，缺少 token_ids 参数。

关键符号：未识别

关键源码片段

vllm/entrypoints/openai/chat_completion/batch_serving.py

批量服务中 parser.parse() 调用点，缺少 token_ids 参数。

```
# vllm/entrypoints/openai/chat_completion/batch_serving.py
# 变更位于 parser.parse() 调用处：增加 model_output_token_ids 参数
if parser is not None:
    reasoning, content, _ = parser.parse(
        output.text,
        request=request, # type: ignore[arg-type]
        model_output_token_ids=output.token_ids, # 新增：传递 token IDs
    )
```

vllm/entrypoints/openai/responses/context.py

Responses API 上下文中的解析调用点，缺少 token_ids 参数。

```
# vllm/entrypoints/openai/responses/context.py
# 变更位于 append_output() 方法中的 parser.parse() 调用
if self.response_parser is not None:
    reasoning, content, tool_calls = self.response_parser.parse(
        completion.text,
        self.request,
        enable_auto_tools=self.enable_auto_tools,
        model_output_token_ids=completion.token_ids, # 新增：传递 token IDs
    )
```

vllm/entrypoints/openai/responses/serving.py

Responses API 输出构造中的解析调用点，缺少 token_ids 参数。

```
# vllm/entrypoints/openai/responses/serving.py
# 变更位于 _make_response_output_items() 方法中的 parser.parse() 调用
if parser:
    reasoning, content, tool_calls = parser.parse(
        final_output.text,
        request,
        enable_auto_tools=self.enable_auto_tools,
        model_output_token_ids=final_output.token_ids, # 新增：传递 token IDs
    )
```

评论区精华

chaunceyjiang 回复 "lgtm"，sfeng33 批准，无实质性讨论或争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。每个调用点仅新增一个关键字参数，且该参数在 chat completion 路径已存在，解析器实现已兼容；未触发任何回归测试失败。

- 影响：影响范围限于 Responses API 和批量服务的非流式解析路径，使这三个入口的 token_ids 传递行为与 chat completion 对齐。对用户无可见行为变化，对后续解析器重构提供基础。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR