PR #38663 完整报告

vllm-project/vllm

[Feat][Core] safely abort requests when FSM fails to advance

合并时间：2026-04-06 23:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38663

执行摘要

- 一句话：修复结构化输出 FSM 失败时请求挂起的 bug，安全中止请求。
- 推荐动作：建议工程师阅读以了解 FSM 失败处理的设计决策，特别是 resumable 字段的重用和状态管理；关注调度器 update_from_output 方法的变更，这对理解结构化输出错误处理有价值。

功能与动机

PR body 指出：'When using structured outputs with the xgrammar backend, streaming requests would hang indefinitely if the FSM failed to advance.' 作者在 Issue 评论中表示在 Cohere 遇到此问题，使得区分程序 bug 和生成超时困难，目的是使失败场景更明确。

实现拆解

主要修改在 vllm/v1/core/sched/scheduler.py 的 update_from_output 方法：添加逻辑检查 FSM 是否拒绝 token，若拒绝则设置 request.status = RequestStatus.FINISHED_ERROR 和 request.resumable = False。移除了原 PR 中的 fsm_failed_to_advance 字段，重用现有 resumable 机制。测试文件添加了两个单元测试验证同步和异步调度器行为。

关键文件：

- vllm/v1/core/sched/scheduler.py（模块 scheduler）：核心调度逻辑修改，添加 FSM 失败检测和处理，确保请求正确中止。
- tests/v1/core/test_scheduler.py（模块 test）：添加同步调度器测试用例，验证 FSM 失败时请求中止行为。
- tests/v1/core/test_async_scheduler.py（模块 test）：添加异步调度器测试用例，覆盖不同调度器场景。

关键符号：update_from_output, test_abort_request_when_structured_output_fsm_cannot_advance

评论区精华

review 中，gemini-code-assist[bot] 指出代码冗余访问 request.structured_output_request，建议使用局部变量；yewentao256 询问拒绝 token 是否传递给下游，作者解释 token 不会传递，请求终止；njhill 建议状态使用 FINISHED_ERROR 而非 ABORTED，并质疑新增字段

必要性，最终作者采纳，简化代码，重用 resumable 字段。所有讨论已解决。

- 代码冗余访问优化 (style): 作者可能已采纳建议，最终代码使用局部变量，但未明确显示在最终 patch 中；从讨论上下文看，问题已解决。
- token 传递下游的正确性 (correctness): 作者 walterbm 解释拒绝的 token 不会传递，请求会终止，确保正确性。
- 状态设置和字段简化设计 (design): 作者采纳建议，移除新字段，在 update_from_output 中设置 request.resumable = False，简化代码并保持一致性。

风险与影响

- 风险：修改了核心调度器逻辑，可能引入回归，影响结构化输出请求的处理；需确保 FSM 失败检测在所有结构化输出后端（如 xgrammar）中正确工作；测试覆盖了特定场景，但未覆盖所有边界情况，如并发请求或复杂语法。
- 影响：用户影响：修复了请求挂起问题，提升流式请求的可靠性和用户体验；系统影响：确保资源正确释放，避免内存或 KV 缓存泄露；团队影响：代码更简洁，使用现有机制，便于维护。
- 风险标记：核心调度器变更，结构化输出边界情况

关联脉络

- PR #38150 [Mistral Grammar] Support Grammar Factory: 涉及结构化输出和语法支持，与本 PR 的 FSM 失败处理相关，都属于结构化输出功能演进。