PR #35076 完整报告

vllm-project/vllm

[Bugfix] Propagate default stop_token_ids to per-request SamplingParams

合并时间：2026-06-30 12:10

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/35076

执行摘要

- 一句话：修复 gpt-oss 工具调用时 stop_token_ids 未传播导致崩溃
- 推荐动作：该 PR 值得精读，特别是协议层参数传播的设计模式：如何从默认参数回退到自定义参数的通用模式。修复虽然简单，但揭示了系统中参数传播的一致性缺陷。对于类似的参数（如 stop、bad_words 等）可参考此模式。

功能与动机

当使用 gpt-oss-20b 模型并启用工具调用时，请求返回 500 错误：HarmonyError: Unexpected token 12606 while expecting start token 200006。原因是模型生成了工具调用 stop token（</call>）后继续生成无关内容，导致 Harmony 解析器崩溃。该问题由 Issue #22519 报告。

实现拆解

1. 在 ChatCompletionRequest 和 CompletionRequest 的 to_sampling_params 中添加 stop_token_ids 合并逻辑 - 文件：vllm/entrypoints/openai/chat_completion/protocol.py 和 vllm/entrypoints/openai/completion/protocol.py - 方法：to_sampling_params() - 在现有的默认参数回退块之后，新增一段代码：先获取请求中的 stop_token_ids，然后检查 default_sampling_params 中是否有 stop_token_ids，如果有则进行合并。如果请求未指定，则直接使用默认值；如果请求已指定，则合并并去重（使用 dict.fromkeys 保持插入顺序）。最后将合并后的 stop_token_ids 传给 SamplingParams.from_optional()。
2. 修改 return 语句中的参数名 - 将原来的 stop_token_ids=self.stop_token_ids 改为 stop_token_ids=stop_token_ids（局部变量），以使用合并后的值。
3. 新增单元测试 - 文件：tests/entrypoints/openai/test_stop_token_ids.py（全新文件，160 行） - 测试类 TestChatCompletionStopTokenIds 和 TestCompletionStopTokenIds，覆盖五种场景：默认值应用、客户端与默认值合并、两者都无、仅客户端、去重。
4. 后续修复（review 中被 cherry-pick） - 移除了最初在 parser/harmony_utils.py 中添加的早期提前 break 逻辑，因为该逻辑会截断含多个工具调用的输出。同时将合并时的 set() 去重改为 dict.fromkeys() 以保持顺序。

关键文件：

- vllm/entrypoints/openai/chat_completion/protocol.py（模块 入口；类别 source；类型 core-logic；符号 ChatCompletionRequest.to_sampling_params）：核心修改文件，在

to_sampling_params 中添加 stop_token_ids 合并逻辑，修复了默认值丢失的 bug。

- vllm/entrypoints/openai/completion/protocol.py（模块 入口；类别 source；类型 core-logic；符号 CompletionRequest.to_sampling_params）：与 chat_completion 相同的修改，保持 CompletionRequest 的行为一致。
- tests/entrypoints/openai/test_stop_token_ids.py（模块 测试；类别 test；类型 test-coverage；符号 TestChatCompletionStopTokenIds, TestCompletionStopTokenIds, test_default_stop_token_ids_applied, test_client_stop_token_ids_merged_with_defaults）：新增完整测试套件，覆盖所有 stop_token_ids 合并场景，确保修复正确且不退化。

关键符号：ChatCompletionRequest.to_sampling_params, CompletionRequest.to_sampling_params

关键源码片段

vllm/entrypoints/openai/chat_completion/protocol.py

核心修改文件，在 to_sampling_params 中添加 stop_token_ids 合并逻辑，修复了默认值丢失的 bug。

```
# vllm/entrypoints/openai/chat_completion/protocol.py
# 在 to_sampling_params() 中，其他默认参数回退之后插入的合并逻辑：

# Merge server-default stop_token_ids (e.g., model-specific tokens
# like </call> for gpt-oss) with any request-specified ones
stop_token_ids = self.stop_token_ids
default_stop_ids = default_sampling_params.get("stop_token_ids")
if default_stop_ids:
    if not stop_token_ids:
        # 客户端未指定时，直接使用服务器默认值
        stop_token_ids = list(default_stop_ids)
    else:
        # 合并并去重，dict.fromkeys 保持插入顺序（客户端 ID 在前）
        stop_token_ids = list(
            dict.fromkeys([*stop_token_ids, *default_stop_ids])
        )

# 后续在调用 SamplingParams.from_optional 时传入这个局部变量
# 原代码直接使用 self.stop_token_ids，现在改用合并后的 stop_token_ids
```

评论区精华

- qandrew要求补充更详细的测试计划，包括服务器启动命令、客户端请求、前后输出。作者随后更新了 PR 描述，包含了完整的复现步骤和结果对比。
- eeerice发现 CI 测试失败，指出早期 break 逻辑会截断多工具调用的输出，提供了修复（使用 dict.fromkeys 保持顺序，并移除 early break）。该修复被 cherry-pick 到本 PR。
- ashishpatel26提交了 follow-up PR，进一步优化顺序保证和 parser null 安全。

- DarkLight1337最终合并了 PR。
- 要求补充测试计划 (question): 作者更新了 PR 描述，添加了完整的测试计划和输出示例。
- CI 测试失败与 *early break* 修复 (design): 作者的 *cherry-picked* 了修复，CI 通过。
- follow-up 优化 (design): 未合入本 PR，但作为后续改进方向。

风险与影响

- 风险：
  - 代码重复：两个协议文件中的合并逻辑完全一样，未来如果修改可能遗漏一处，存在维护风险。
  - 顺序依赖：使用 dict.fromkeys 保持顺序，但合并后的列表中客户端指定的 ID 优先（在前），如果有依赖于 token 顺序的模型逻辑可能产生行为变化，但理论上影响不大。
  - 性能：合并操作仅发生在每个请求的 to_sampling_params 中，性能开销可忽略。
  - 安全性：无直接影响。
  - 兼容性：对于未使用工具调用的模型，可能有新的 stop_token_ids 传入 sampler，但 SamplingParams 应能正确处理额外 stop token。
- 影响：
  - 用户：使用 gpt-oss 模型并启用工具调用的用户可以正常使用，不再崩溃。其他模型不受影响。
  - 系统：每请求多一次字典查询和列表合并，性能影响可忽略。
  - 团队：维护者需要关注两个 protocol 文件的同步修改。
  - 风险标记：合并逻辑在双端重复，去重方式依赖 dict 顺序

关联脉络

- 暂无明显关联 PR