PR #38992 完整报告

vllm-project/vllm

[Bugfix] Fix invalid JSON in Gemma 4 streaming tool calls by stripping partial delimiters

合并时间：2026-04-06 01:11

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38992

执行摘要

该 PR 修复了 Gemma 4 模型在流式工具调用中，因令牌边界切分字符串分隔符导致部分分隔符片段（如 '<'、'|'）泄漏到 JSON 参数中，从而引发 JSON 解析错误的 bug。通过扩展参数解析时剥离的字符集，并添加单元测试覆盖此边界情况，确保流式输出的 JSON 始终有效。影响范围限于 Gemma 4 工具调用，风险较低，是重要的可靠性修复。

功能与动机

修复 Issue #38946 中报告的问题：Gemma 4 流式工具调用会生成无效 JSON。问题根源在于，当流式输出中令牌边界切分了 Gemma 4 的字符串分隔符 <|"|> 时，残留的分隔符片段（如 '<'、'|'）未被剥离，混入 JSON 参数值中，导致后续解析失败。PR body 明确引用该 issue，并指出“Partial <|"|> delimiter chars must not leak into streamed JSON”。

实现拆解

主要改动涉及两个文件：

1. 核心解析逻辑（vllm/tool_parsers/gemma4_tool_parser.py）：
   - 在 _emit_argument_diff 方法中，修改 safe_json 的生成逻辑，将剥离的字符从 "}"、'"'、"]" 扩展为还包括 "<"、"|"、"\\"、">"。
   - 代码片段：
   - 这确保了即使分隔符被部分流式输出，其片段也不会污染 JSON。
2. 单元测试（tests/tool_parsers/test_gemma4_tool_parser.py）：
   - 新增 test_streaming_split_delimiter_no_invalid_json 测试，模拟分隔符被切分的流式 chunks：
   - 验证参数文本可被 json.loads 解析，且不包含 '<' 等分隔符片段。

评论区精华

review 讨论简洁，主要聚焦于修复验证：

- bbrowning：

“I cloned this locally, verified the new unit test reproduces the failure without the parser change and then verified the parser change fixes the unit test. The change looks good to me and the additional withheld characters here match what the

official Gemma 4 prompt formatting docs show as the string delimiters."

- 确认了修复的有效性，并指出剥离字符集与官方文档一致，支持了设计合理性。
- 其他 reviewer（robertgshaw2-redhat、tlrmchlsmth）无具体评论，直接批准。无争议或未解决疑虑。

风险与影响

- 风险：
  - 剥离字符集扩展可能过度剥离有效字符（如 JSON 字符串中合法出现的 '<'、'>'），但 Gemma 4 的 JSON 参数通常不包含这些字符，风险可控；现有测试应覆盖常规场景。
  - 性能影响可忽略，仅增加少量字符检查。
- 影响：
  - 用户：修复后，Gemma 4 流式工具调用用户不再遇到 JSON 解析错误，提升可靠性。
  - 系统：仅影响工具解析模块，不涉及核心推理路径或其他模型。
  - 团队：新增测试作为回归防护，变更简单易维护。

关联脉络

- 从近期历史 PR 看，本 PR 与工具调用（tool-calling）相关，但未发现直接关联的 PR。
- 作为 bugfix，它独立解决了 Gemma 4 模型在流式场景下的一个边界问题，体现了对模型特定格式处理的精细化改进。