PR #38909 完整报告

vllm-project/vllm

[Bugfix][Frontend] Fix Gemma4 streaming HTML duplication after tool calls

合并时间：2026-04-08 11:03

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38909

执行摘要

- 一句话：修复 Gemma4 流式工具解析器中 HTML 内容重复的 bug。
- 推荐动作：对于处理工具解析或 Gemma4 模型的工程师值得精读，学习缓冲区管理在流式解析中的正确实践，并参考新增测试作为回归防护示例。

功能与动机

根据 Issue #38910 描述，Gemma4 工具解析器在流式模式下会损坏普通文本，如将 变成 < ，或在 HTML 工具参数中重复标签前缀如 <<htmlhtml。PR body 指出根因是 current_text 被错误重建，导致无效中间状态。

实现拆解

主要改动在 vllm/tool_parsers/gemma4_tool_parser.py 的 extract_tool_calls_streaming 方法中，移除了 current_text = previous_text + delta_text 行，改为直接使用上游的 current_text，避免缓冲 delta 污染累积文本。测试文件 tests/tool_parsers/test_gemma4_tool_parser.py 添加了两个回归测试：test_streaming_does_not_duplicate_plain_text_after_tool_call 和 test_streaming_html_argument_does_not_duplicate_tag_prefixes，覆盖普通文本和 HTML 参数场景。

关键文件：

- vllm/tool_parsers/gemma4_tool_parser.py（模块 tool_parsers）：修复核心解析逻辑，移除错误的重建 current_text 代码，确保流式缓冲区正确处理。
- tests/tool_parsers/test_gemma4_tool_parser.py（模块 test）：添加回归测试，验证普通文本和 HTML 参数不被重复，提供关键防护。

关键符号：Gemma4ToolParser.extract_tool_calls_streaming

评论区精华

Review 评论中没有技术争议，所有 review 者（gemini-code-assist[bot]、sfeng33、chaunceyjiang）都批准了 PR。gemini-code-assist[bot] 分析确认变更解决重复问题，Issue 评论中仅涉及合并冲突解决，无设计权衡讨论。结论是修复被接受并验证有效。

- 修复验证 (correctness): PR 被接受并合并，无争议。

风险与影响

- 风险：风险较低：变更仅限于 Gemma4 工具解析器的流式路径，移除错误逻辑可能引入回归，但新增测试覆盖了关键场景。需注意其他解析器可能不受影响，但 PR 范围集中，且通过测试验证。
- 影响：影响使用 Gemma4 模型进行流式工具调用的用户，修复了输出损坏问题，提升解析正确性和用户体验。对系统其他部分无直接影响，仅限于前端工具解析模块。
- 风险标记：流式解析错误，回归测试添加

关联脉络

- PR #38848 [Bugfix] Fix Qwen3 tool parser for Responses API tools: 同属工具解析器 bugfix，涉及相似模块和测试覆盖。
- PR #38860 [Parser] Pass request.tools to tool parser: 涉及前端工具解析逻辑修复，显示工具调用解析的持续改进。