PR #20004 完整报告

sgl-project/sglang

Multi tool streaming fix

合并时间：2026-04-02 12:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20004

执行摘要

本 PR 修复了 Qwen25 模型在流式多工具调用解析中的失败问题，通过修改解析器添加回退机制，确保所有工具调用能正确解析，提升了功能一致性和用户体验。

功能与动机

该 PR 旨在解决 issue 18102 中报告的问题：在使用 Qwen3-Next-80B-A3B-Thinking 模型进行 speculative decoding 时，流式多工具调用解析失败，仅返回第一个工具调用。PR body 明确指出修复目标为块状格式（如 Qwen25）的流式多工具调用解析，避免因非 JSON 标记导致的解析错误。

实现拆解

主要改动集中在两个文件：

1. python/sglang/srt/function_call/base_format_detector.py：修改了 parse_streaming_increment 方法。 - 添加 used_separator_branch 标志以跟踪解析分支。 - 当 JSON 解析失败时，回退到搜索 bot_token 跳过非 JSON 标记（例如 Qwen25 的 </tool_call>\n<tool_call>\n）。 - 扩展异常处理从仅捕获 MalformedJSON 到捕获 MalformedJSON 和 json.JSONDecodeError。
2. test/registered/unit/function_call/test_function_call_parser.py：新增 TestQwen25Detector 类，包含 7 个测试用例，覆盖单工具调用、多工具调用的流式和非流式场景，确保修复有效性。

评论区精华

Review 讨论较少，两位 reviewer（JustinTong0323 和 hnyls2002）均快速批准了 PR，没有深入讨论或争议。这表明变更被认为直接、必要且风险较低。

风险与影响

- 风险：修改了解析核心路径，可能影响其他格式模型的解析行为，引入回归；异常处理逻辑变化可能掩盖潜在错误。但新增测试覆盖降低了这些风险。
- 影响：主要影响使用 Qwen25 模型的用户，修复后流式多工具调用解析将正确工作，提升功能一致性；对系统其他部分无直接影响，但增强了解析模块的健壮性。

关联脉络

从历史 PR 看，PR 21225 涉及 speculative decoding 和解析改进，与本 PR 的 issue 场景相关（issue 18102 提到 speculative decoding）。这表明项目在持续优化解析和推测解码功能，本 PR 是这一演进方向中的一环，专注于工具调用解析的细节修复。