PR #37386 完整报告

vllm-project/vllm

fix(glm47): improve tool call parsing and content normalization

合并时间：2026-03-18 16:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37386

执行摘要

本 PR 修复了 GLM-4.7 模型工具调用解析的 bug，通过改进正则表达式以正确处理函数名和参数捕获，并将空内容规范化到 None 以符合 OpenAI API 标准。影响范围限于 GLM-4.7 和 GLM-4.5 模型的工具调用模块，提升了稳定性和兼容性。

功能与动机

本 PR 旨在解决 Issue #37277 报告的 GLM-4.7 工具调用失败问题。PR body 中明确表示：“需要改进 GLM-4.7 解析逻辑以处理格式差异（如函数名后无换行符、零参数调用），并规范内容为空时的返回值为 None，以遵循 OpenAI API 约定。”关联的 Issue #32436 和 #33877 也涉及类似解析错误，表明这是一个持续性 bug。

实现拆解

关键改动点如下：

- 正则表达式优化：在 vllm/tool_parsers/glm47_moe_tool_parser.py 中，func_detail_regex 从 r"<tool_call>(.*?)(<arg_key>.*?)?</tool_call>" 改为 r"<tool_call>\s*(\S+?)\s*(<arg_key>.*)?</tool_call>"，使用 \S+? 避免尾部空白，func_arg_regex 从 r"<arg_key>(.*?)</arg_key>(?:\\n|\s)*<arg_value>(.*?)</arg_value>" 简化为 r"<arg_key>(.*?)</arg_key>\s*<arg_value>(.*?)</arg_value>"。
- 内容规范化：在 vllm/tool_parsers/glm4_moe_tool_parser.py 的 *extract_tool_calls* 方法中，添加逻辑：如果 *content* 为空或仅空白，则设置为 None。
- 测试覆盖：新增 tests/tool_parsers/test_glm47_moe_tool_parser.py 文件，包含测试用例如零参数调用、内联参数、换行参数等；更新 tests/tool_parsers/test_glm4_moe_tool_parser.py 中的 *expected_content* 从 "" 改为 None。

评论区精华

Review 讨论有限，主要结论为整体批准。*gemini-code-assist[bot]* 评论：“改进增强了正确性和可维护性”，*chaunceyjiang* 回复“LGTM.”。Issue 评论中，用户 xi1212 报告了类似问题，但后来澄清是 tool_choice 配置错误，提示变更可能对用户环境敏感。

风险与影响

风险：正则表达式变更可能引入新的解析错误，例如处理嵌套标签时；内容类型从空字符串改为 None 可能破坏依赖空字符串的下游代码；测试虽全面，但未覆盖所有可能的模型输出变体。
影响：对用户，修复了工具调用失败，提升体验；对系统，解析更健壮，减少错误；对团队，新增测试有助于维护，但需关注内容类型变更的连锁反应。

关联脉络

本 PR 直接关联 Issue #37277、#32436 和 #33877，表明 GLM 模型工具调用解析是一个反复出现的问题。近期历史 PR 中无直接相关项，但工具调用模块的 bugfix 显示 vLLM 在持续优化模型兼容性和 API 标准化。