

PR #33898 完整报告

sgl-project/sglang

[inking] Render tool-result media instead of coercing content to str

合并时间: 2026-08-08 14:39

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33898>

执行摘要

- 一句话: 修复 Inkling tool 结果渲染图片报 400 错误
- 推荐动作: 值得精读。核心设计点是 "渲染器输出与 MM processor 媒体收割之间的占位符计数契约", 这类隐式契约在推理服务里很容易被忽略, 本 PR 用专门测试把它显式固化; 同时展示了如何把特判分支收敛为复用统一解析逻辑。建议关注后续 Inkling 多模态链路 (如 audio part) 是否也有类似契约需要补齐。

功能与动机

PR body 明确指出: Multimedia (image) tool results did not render correctly in Inkling's OSS renderer, leading to 400 Bad Request errors。提交信息进一步给出根因: render_inking_messages 对 tool 角色特判, 把 content 塞进 _expect_string_content, 该函数对非 str 内容一律抛 TypeError; 而 Anthropic 适配器的 _convert_tool_result_content 对任何不止一个文本块的 tool_result 都会产出 list, 于是携带图片的 tool 结果 (Claude Code 截图、读回渲染文件) 让整个请求失败: TypeError: message content must be a string for this Inkling role, got list。

实现拆解

1. 变更入口: render_inking_messages (python/sglang/srt/parser/inking_renderer.py) 主循环中的 if role == "tool" 分支。
2. 核心逻辑改造: 将 _expect_string_content(message.get("content", "")) 替换为 list(_iter_render_parts(...)) 遍历, 逐 part 调用 _append_message, 使 text、image、audio 都按自身类型渲染, 并在 tool 分支内显式拒绝 thinking part, 与通用分支语义对齐。
3. 边界处理: 对空 content ("" / None / []) 回退为 [("text", "")], 避免已应答的 tool_call 悬空; tool_name 用 str() 包裹, 保证 author_name 始终为字符串。
4. 配套清理: 删除不再被引用的 _expect_string_content 辅助函数。
5. 测试配套: test/registered/unit/parser/test_inking_renderer.py 新增 6 个用例, 覆盖图片 + 文本混合渲染、占位符计数与图片数一致、多文本块逐一渲染、空结果仍产出块、author 回退到 tool_call_id、拒绝 thinking part。
6. 验证配套: 作者手工回放 106 条采集的 agent 请求 (含 14 条图片、5 条多文本 tool 结果), 修复前 19 条抛错, 修复后 0 条, 且无占位符 / 媒体数量错配。

关键文件:

- python/sglang/srt/parser/inkling_renderer.py (模块 渲染器; 类别 source; 类型 core-logic; 符号 render_inkling_messages, _iter_render_parts, _expect_string_content) : 核心修复文件: tool 分支从字符串强转改为逐 part 渲染, 删除 _expect_string_content, 是本 PR 的行为变更主体。
- test/registered/unit/parser/test_inkling_renderer.py (模块 渲染测试; 类别 test; 类型 test-coverage; 符号 test_tool_result_renders_image_parts_alongside_text, test_tool_result_placeholder_count_matches_image_parts, test_tool_result_with_multiple_text_blocks_renders_each, test_empty_tool_result_still_emits_a_block) : 新增 6 个回归测试, 明确锁定图片占位符计数契约与空结果边界行为, 是本次修复质量的关键保障。

关键符号: render_inkling_messages, _iter_render_parts, _append_message, _expect_string_content

关键源码片段

python/sglang/srt/parser/inkling_renderer.py

核心修复文件: tool 分支从字符串强转改为逐 part 渲染, 删除 _expect_string_content, 是本 PR 的行为变更主体。

```
# render_inkling_messages 主循环中处理 tool 角色的分支 (合并后版本)
for message_index, message in enumerate(message_list):
    if message_index == leading_system_count:
        append_effort()
    role = _expect_role(message)
    if role == "tool":
        # 修复前这里用 _expect_string_content 把 content 强转成 str,
        # 而 Anthropic 适配器对非单文本块的 tool_result 产出 list,
        # 带图片的 tool 结果 (agent 截图、读回 PNG) 会抛 TypeError 导致请求 400。
        # 现在复用 _iter_render_parts 逐 part 渲染, MM processor 会按媒体项数量展开占位符,
        # 两者必须保持一致, 否则渲染与收割两趟会错位 (desync) 。
        tool_name = str(
            message.get("name")
            or tool_call_id_to_name.get(message.get("tool_call_id") or "", "")
        )
        tool_parts = list(_iter_render_parts(message.get("content", "")))
        if not tool_parts:
            # 空 tool 结果也要产出一个空 text 块, 否则它对应的 tool_call 会悬空
            tool_parts = [("text", "")]
        for kind, text in tool_parts:
            # thinking part 只允许出现在 assistant 角色, tool 角色出现直接报错
            if kind == "thinking":
                raise ValueError("Inkling thinking parts require role='assistant'")
            _append_message(
                input_ids,
                tokenizer,
                "tool",
                kind,
```

```
text,
author_name=tool_name,
)
continue
```

评论区精华

本次合并没有行内 review 评论，讨论集中在流程与验证：yhyang201 在评论中 @JustinTong0323 表示 LGTM，随后两位 reviewer 均 APPROVED；CI 曾触发重跑，ispobock 用 /rerun-test 单独跑 test/registered/unit/parser/test_inkling_renderer.py，github-actions 机器人回报 1 个测试全部通过。PR body 中的验证数据（106 条请求回放，19 条报错降到 0 条）是合入的重要依据。

- 106 条真实 agent 请求回放验证 (testing): 作为合入依据的验证数据，说明修复覆盖了真实场景，而非仅单元测试通过。
- Review 审批与 LGTM (other): 双 approve 后合入 main，无设计层面的争议。
- CI 重跑与渲染器单测验证 (other): 测试通过后 PR 合入，属于常规 CI 流程。

风险与影响

- 风险：
 1. 渲染路径行为变更：tool 内容从 " 要么 str 要么抛错 " 变为 " 逐 part 渲染 "，依赖旧报错语义的下游逻辑（概率极低）会受影响。
 2. 占位符计数隐式契约：渲染器与 MM processor 之间约定 IMAGE_TOKEN_ID 数量必须等于媒体收割数量，本 PR 用 test_tool_result_placeholder_count_matches_image_parts 锁定该契约，但 MM processor 侧收割逻辑后续若有改动，需同步回归。
 3. 手工回放验证未自动化：106 条请求重放是作者手工完成，仓库中没有对应的自动化端到端回归用例，后续回归只能靠单元测试部分覆盖。
 4. 影响范围有限：仅改 tool 分支，user/assistant 路径与推理 kernel 无改动，无性能影响。
 - 影响：修复 Inkling OSS 渲染器在 tool-use/agent 场景下的多模态渲染缺陷，携带图片的 tool 结果从 400/500 变为正常渲染，影响 Claude Code 风格 agent 工作流（截图、读回文件等）。影响范围为渲染器 tool 分支及其与 MM processor 的交互，对用户/assistant 消息处理无影响；团队侧收益是可复用的 _iter_render_parts 统一了 tool 与普通角色的 part 解析逻辑，删除的死代码降低了维护成本。
 - 风险标记：渲染分支行为变更，媒体占位符计数契约，手工回放验证未自动化

关联脉络

- PR #33417 Fix deterministic inference for Inkling: 同为 Inkling 支持链路的修复：33417 解决推理确定性 /logprob 不一致，本 PR 解决渲染层 tool 结果报错，两者都围绕 Inkling 模型在 sglang 中的可用性。
- PR #33903 [Inkling] silu_and_mul: replace helion kernels with plain Triton: 同为 Inkling 模块的演进（MoE 内核迁移），与本 PR 的渲染层改动共同构成 Inkling 完整支持脉络。

- PR #34009 Add the 8-gpu Inkling consistency test: 持续完善 Inkling 测试覆盖，与本 PR 的渲染回归测试互补，说明 Inkling 在仓库中处于快速建设期。