

# PR #34781 完整报告

sgl-project/sglang

fix(muse-glimmer): parse required/named tool calls natively

合并时间: 2026-08-15 04:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34781>

## 执行摘要

- 一句话: MuseGlimmer required/named 工具调用改走原生 ATEM 解析
- 推荐动作: 值得快速阅读: PR 体小但清晰解释了 SGLang 工具调用解析器的路由机制 (JsonArrayParser 与 detector 原生解析的切换条件)。重点关注 `parses_required_natively` 这一行为开关的设计: 它以方法形式暴露「模型是否原生产出工具调用格式」的语义, 比在路由层硬编码模型名更易扩展。建议后续补一个多轮 required/named 的回归测试, 并在代码注释或文档中说明该开关与 JSON-array grammar 路径的关系。

## 功能与动机

PR body 明确指出: Muse Glimmer 只产出 ATEM () 工具调用, 当 `tool_choice="required"` 或指定命名工具时, 默认路径路由到 JsonArrayParser 并用 JSON-array grammar 约束生成; 模型不产出该 JSON 格式, 导致调用泄漏为 content、`finish_reason="stop"` 且零 `tool_calls`。该问题在多轮对话 (历史已含 assistant `tool_calls` 轮次) 下可稳定复现, 影响 required/named 工具调用的核心功能。

## 实现拆解

1. 问题定位: 在 `python/sglang/srt/function_call/muse_glimmer_detector.py` 中确认 MuseGlimmerDetector 已具备原生 ATEM 解析能力 (`tool_choice=auto` 路径可用), 但 required/named 路径被框架路由到 JsonArrayParser, 解析器与模型输出格式不匹配。
2. 核心改动: 在 MuseGlimmerDetector 中新增方法 `parses_required_natively()` 并返回 True。该方法是一个行为开关, 框架据此将 required/named `tool_choice` 从 JSON-array grammar 路径切换到 detector 的原生 ATEM 解析路径, 与 auto 行为对齐。
3. 配套验证: PR body 给出 Muse-Glimmer-30B (BF16 单卡) 在 main 上的前后对比, required/named 多轮对话场景从「0 tool calls、ATEM 泄漏为 content」修复为正常返回工具调用; 现有 `test/registered/unit/function_call/` 测试套件通过; 未新增回归测试, 无配置、文档或部署配套改动。

关键文件:

- `python/sglang/srt/function_call/muse_glimmer_detector.py` (模块 函数调用; 类别 source; 类型 core-logic; 符号 `parses_required_natively`): 核心修复文件: 为 MuseGlimmerDetector 新增 `parses_required_natively()` 并返回 True, 使

required/named tool\_choice 跳过 JsonArrayParser 的 JSON-array grammar, 改走原生 ATEM 解析路径, 修复多轮对话下工具调用丢失问题。

关键符号: parses\_required\_natively

## 关键源码片段

### python/sglang/srt/function\_call/muse\_glimmer\_detector.py

核心修复文件: 为 MuseGlimmerDetector 新增 parses\_required\_natively() 并返回 True, 使 required/named tool\_choice 跳过 JsonArrayParser 的 JSON-array grammar, 改走原生 ATEM 解析路径, 修复多轮对话下工具调用丢失问题。

```
# python/sglang/srt/function_call/muse_glimmer_detector.py
# 关键约定: Muse Glimmer 只通过 ATEM (<atem:function_calls>) 区块表达工具调用,
# 不产出 JSON 数组格式; 以下方法共同定义 parser 的行为边界。
```

```
def supports_structural_tag(self) -> bool:
    # ATEM 块不使用结构化 tag 注入, 因此固定返回 False
    return False
```

```
def parses_required_natively(self) -> bool:
    """声明 required/named tool_choice 可由本 detector 原生解析。
```

```
    默认路径在 required 或命名工具时会通过 JsonArrayParser 强制
    JSON-array grammar, 但模型实际输出原生 ATEM 块, JSON 路径无法
    解析, 导致调用泄漏为 content 且 tool_calls 为空。返回 True 后,
    required/named 与 auto 共用同一原生 ATEM 解析路径。
    """
```

```
    return True
```

```
def structure_info(self) -> _GetInfoFunc:
    # 为解码器提供 ATEM 结构的拼接信息, begin/end 分别对应 open/close 标签
    return lambda name: StructureInfo(
        begin=f'{FUNCTION_CALLS_OPEN}\n{INVOKE_OPEN} name="{name}">',
        end=f'{INVOKE_CLOSE}\n{FUNCTION_CALLS_CLOSE}',
        trigger=FUNCTION_CALLS_OPEN,
    )
```

## 评论区精华

review过程简洁: JustinTong0323审阅后直接批准 ("LGTM"), 未提出代码层面的修改意见。第二个 commit 由 JustinTong0323 将 parses\_required\_natively 的多行 docstring 修剪为更精炼的版本, 说明评审关注注释简洁性。合并前维护者通过 /tag-and-rerun-ci 与 /rerun-failed-ci 重跑 CI, 其中 base 常规运行失败、extra 运行通过。

- 整体评审结论 (other): APPROVED, 改动可直接合并。
- docstring 精简 (style): 已通过后续 commit 修正。
- CI 重跑与结果不一致 (other): extra CI 通过, base 失败原因未在讨论中明确澄清。

## 风险与影响

- 风险:

1. 行为开关风险: `parses_required_natively()` 返回 `True` 后, Muse Glimmer 的 `required/named` 路径整体切换解析器。若未来模型更新开始产出 JSON-array 格式, 该开关将导致此类输出无法被解析; 目前模型仅产出 ATOM, 风险低, 但缺少对开关语义的显式约束。
2. 测试覆盖缺口: PR 未新增回归测试, 多轮 `required/named` 场景的回归保护完全依赖现有 `function_call` 套件, 后续重构 `JsonArrayParser` 路由时可能重新引入该 bug 而无法被及时发现。
3. CI 不一致: base 常规 CI 运行失败而 extra 运行通过, 虽未发现与改动相关的失败证据, 但合并前未明确澄清失败原因, 存在轻微不确定性。 - 影响: 影响范围集中在 Muse-Glimmer 模型的工具调用解析: 使用 `tool_choice=required` 或命名工具的用户, 尤其是多轮对话场景, 工具调用不再丢失, 修复前功能完全不可用、修复后与 `auto` 行为一致。系统层面推理性能影响, 仅改变已有输出的解析路径选择; 单文件 5 行改动, 对其他模型和解析器无影响。团队层面评审成本低, 但建议后续补齐回归测试。 - 风险标记: 缺少新增回归测试, `required/named` 解析行为整体切换, CI base 运行失败未明确归因

## 关联脉络

- 暂无明显关联 PR