PR #20208 完整报告

sgl-project/sglang

Remove maxItems=1 restriction when tool_choice is specified

合并时间：2026-04-03 10:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20208

执行摘要

该 PR 修复了函数调用 JSON 模式中 maxItems=1 限制导致的模型停滞问题，通过引入 parallel_tool_calls 参数控制是否限制单次调用，默认允许并行调用，与 OpenAI API 兼容，修复了 Issue #17998 中的 bug，提升函数调用可靠性。

功能与动机

修复 Issue #17998 中描述的 bug：当 tool_choice 指定函数时，JSON 模式强制 maxItems: 1，导致模型在提示暗示多次调用时停滞生成空白字符。PR body 明确引用此 Issue 作为动机，旨在解除不必要的限制，支持并行工具调用场景。

实现拆解

关键改动点按模块梳理：

- function_call 模块：在 get_json_schema_constraint 函数中移除无条件 maxItems: 1，添加 parallel_tool_calls 参数，仅在 False 时设置限制；在 get_structure_constraint 函数中传递该参数。
- openai API 模块：在 ChatCompletionRequest 类添加 parallel_tool_calls: bool = True 字段，扩展 API 兼容性；在 serving_chat.py 中处理参数传递。
- 测试模块：移除不稳定集成测试 test_specific_function_multi_call_no_stall，添加单元测试覆盖 parallel_tool_calls=True 和 False 的行为，验证 JSON 模式正确性。

评论区精华

review 讨论中仅 JustinTong0323 批准 PR；但 Issue 评论揭示关键交锋：hnyls2002 报告集成测试失败，指出 parallel_tool_calls 默认值问题，导致 kpham-sgl 纠正假设并更新实现。引用 hnyls2002 原话："test_specific_function_multi_call_no_stall failed..."，以及 kpham-sgl 回应："oh I assumed parallel_tool_calls is already default to True... Good catch!"，体现了测试驱动修复的过程。

风险与影响

风险：新增参数默认值可能影响现有行为，但设为 True 保持向后兼容；JSON 模式修改需确保无回归错误，已有单元测试覆盖；新参数需文档说明以避免用户混淆。 影响：修复 bug 提升用户函数调用体验，支持并行调用扩展功能；系统层面 API 对齐 OpenAI 标准，无性能影响；团

队需注意新参数的集成和测试维护。

关联脉络

从历史 PR 分析，近期 PR 如 #21570（支持 LoRA）和 #19163（Transformers 后端）涉及模型功能扩展，但本 PR 专注于函数调用 JSON 模式修复，属于独立 bugfix。与 Issue #17998 直接关联，展示了针对具体问题的快速响应和设计权衡（通过参数控制限制）。整体看，SGLang 在持续优化 API 兼容性和功能可靠性。