PR #22175 完整报告

sgl-project/sglang

fix: server crash when stop_token_ids contains null

合并时间：2026-04-11 02:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22175

执行摘要

- 一句话：修复 stop_token_ids 包含 null 时服务器崩溃问题，过滤 None 值防止下游 torch.tensor 异常。
- 推荐动作：该 PR 值得快速浏览以了解防御性输入验证的模式。虽然变更简单，但展示了如何处理 JSON null 值在 Python 中的传播问题。关注点：
  1. 从提交历史看代码如何从复杂实现简化为最终简洁版本。
  2. review 中提到的未修复的类似漏洞（stop 和 stop_regex）值得后续关注。

功能与动机

修复 Issue #22144 中报告的服务器崩溃问题。当客户端发送包含 "stop_token_ids": [null] 的 JSON 请求时，null 值会传播到 torch.tensor() 调用中，导致调度器崩溃。PR body 明确指出需要过滤 SamplingParams.__init__ 中的 None 值。

实现拆解

核心改动在 python/sglang/srt/sampling/sampling_params.py 文件的 __init__ 方法中：

1. 修改 stop_token_ids 处理逻辑，从 set(stop_token_ids) 改为 {int(t) for t in stop_token_ids if t is not None}，过滤 None 值并强制转换为 int。
2. 当过滤后集合为空时，将 self.stop_token_ids 设为 None。
3. 提交历史显示代码经过多次简化：从最初的复杂过滤逻辑简化为最终的单行推导式，并移除了额外的 token ID 范围检查。

关键文件：

- python/sglang/srt/sampling/sampling_params.py（模块 sampling）：唯一修改的文件，修复了 stop_token_ids 中 None 值过滤的核心逻辑。

关键符号：SamplingParams.init

评论区精华

review 中只有 gemini-code-assist[bot] 的一条评论，建议将类似的防御性过滤扩展到 stop 和 stop_regex 参数，因为这些参数同样可能从 JSON 负载接收 null 值，在 normalize() 方法中处理时可能导致崩溃。但 PR 作者没有回应此建议，最终只修复了 stop_token_ids。

- 防御性过滤应扩展到 stop 和 stop_regex 参数 (correctness): PR 作者未回应此建议，最终只修复了 stop_token_ids。

风险与影响

- 风险：风险较低：
  1. 回归风险：修改范围极小（仅 3 行变更），逻辑简单直接，不太可能引入新 bug。
  2. 兼容性：过滤 None 值不会影响正常整数 token ID 的处理，但可能改变行为：原本包含 null 的列表会导致崩溃，现在会被静默忽略。这符合防御性编程原则，但可能掩盖客户端错误。
  3. 未修复的潜在风险：如 review 评论指出，stop 和 stop_regex 参数仍有类似漏洞，可能在未来导致崩溃。
- 影响：影响范围有限但重要：
  1. 用户影响：修复了特定 JSON 输入导致的服务器崩溃，提升了 API 鲁棒性，防止恶意或错误请求导致服务中断。
  2. 系统影响：避免了调度器异常退出，提高了服务稳定性。
  3. 团队影响：这是一个简单的输入验证修复，不涉及核心算法或架构变更，维护成本低。
- 风险标记：输入验证不完整，潜在类似漏洞未修复

关联脉络

- PR #22144 [Bug] "stop_token_ids": [null] causes server to crash: 这是本 PR 要修复的 Issue，直接关联。
- PR #22312 Make GDN support non-continuous B/A Tensor input to fix the accuracy regression of Qwen3.5-27B: 同为 bugfix 标签，涉及输入验证和鲁棒性改进。
- PR #22495 Add page_size to admission token budget check: 同为 bugfix 标签，涉及调度器相关修复。