

PR #31973 完整报告

sgl-project/sglang

Guard min_new_tokens penalizer against None eos_token_id

合并时间: 2026-07-22 23:24

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31973>

执行摘要

- 一句话: 修复 eos_token_id 为 None 时 min_new_tokens 崩溃
- 推荐动作: 建议合入, 属于明确的边界 bug 修复, 代码简洁, 风险低。可快速合入。

功能与动机

修复 Issue #31972: min_new_tokens 惩罚器在 tokenizer.eos_token_id 为 None 时崩溃。原代码将 None 插入 int64 张量, 触发异常。

实现拆解

1. 在 python/sglang/srt/sampling/penaltylib/min_new_tokens.py 的 _prepare 方法中, 将构造 stop token 集合的表达式由 {req.tokenizer.eos_token_id} 改为条件表达式: {req.tokenizer.eos_token_id} if req.tokenizer.eos_token_id is not None else set()。
2. 若无 eos_token_id, 则使用空集合, 避免 None 进入张量。
3. 未修改其他逻辑, 不影响正常情况下的行为。

关键文件:

- python/sglang/srt/sampling/penaltylib/min_new_tokens.py (模块 采样器; 类别 source; 类型 core-logic; 符号 _prepare): 核心修复文件, 增加 eos_token_id 非 None 检查

关键符号: _prepare

关键源码片段

[python/sglang/srt/sampling/penaltylib/min_new_tokens.py](#)

核心修复文件, 增加 eos_token_id 非 None 检查

```
# python/sglang/srt/sampling/penaltylib/min_new_tokens.py
# 修正后的 _prepare 方法片段: 仅在 eos_token_id 不为 None 时加入集合
def _prepare(self):
    # ... 前面的代码省略 ...
    padded_stop_token_ids = torch.nn.utils.rnn.pad_sequence(
        sequences=[
            torch.tensor(
                data=(
                    list(
```

```

        (req.sampling_params.stop_token_ids or set())
        | (req.tokenizer.additional_stop_token_ids or set())
        | (
            {req.tokenizer.eos_token_id}
            if req.tokenizer.eos_token_id is not None # 新增守卫: 避免 None 进入 int64
            张量
            else set() # eos_token_id 为 None 时使用空集合
        )
    ),
    dtype=torch.int64,
    device=self.orchestrator.device,
)
for req in self.orchestrator.reqs()
],
batch_first=True,
padding_value=self.orchestrator.vocab_size,
)
# ... 后续代码不变 ...

```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：改动仅一行条件守卫，不影响 eos_token_id 正常存在时的行为。需注意若其他位置有类似模式也应检查，但当前 PR 范围安全。
- 影响：影响范围：使用 min_new_tokens 或 min_tokens 参数且模型 tokenizer 无 eos_token_id 的用户。修复后请求不再崩溃，min_length 约束通过其他 stop token 执行。对大多数模型无影响。
- 风险标记：暂无

关联脉络

- PR #31975 Treat partial_json_parser AssertionError as incomplete JSON: 同为采样 / 工具调用相关的 bugfix，涉及 tensor 构造时处理 None/ 异常边界情况