

PR #31378 完整报告

sgl-project/sglang

[Bugfix] Fix min-new-token EOS handling

合并时间: 2026-08-20 06:29

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31378>

执行摘要

- 一句话: 修复 min-new-token 对模型配置 EOS 的处理与崩溃
- 推荐动作: 该 PR 值得精读, 尤其关注 `_prepare()` 中如何合并 EOS 集合和过滤 `None` 值。这是一个小而精准的 bugfix, 展示了如何处理 tokenizer 与模型配置不一致的情况。

功能与动机

PR body 指出: 请求可能携带来自模型配置的有效 EOS ID, 即使 tokenizer 未暴露 EOS ID。原来的 min-new-token 惩罚器省略了这些 ID, 并试图把 `None` 转换为整数 tensor, 导致批量准备崩溃。修复后, 采用与正常结束检测相同的请求 EOS 集合, 避免崩溃并确保配置的 EOS token 在达到最小长度前一直被抑制。

实现拆解

1. 修改 `python/sglang/srt/sampling/penaltylib/min_new_tokens.py` 的 `BatchedMinNewTokensPenalizer._prepare()` 方法:
 - 在构建 `stop_token_ids` 集合时, 合并 `req.eos_token_ids` (模型配置的 EOS token ID 集合)
 - 使用列表推导式并在生成 tensor 前过滤掉 `None` 值, 避免 `None` 转换为整数 tensor 时崩溃。
2. 修改测试文件 `test/registered/unit/sampling/test_penaltylib.py`:
 - 在 `_make_req` 辅助函数中增加 `req.eos_token_ids = None` 默认设置。
 - 新增测试 `test_blocks_model_eos_without_tokenizer_eos`, 验证当 tokenizer EOS 缺失时, 模型配置的 EOS token 仍被阻止。
3. 无额外配置或部署改动。

关键文件:

- `python/sglang/srt/sampling/penaltylib/min_new_tokens.py` (模块 采样惩罚; 类别 source; 类型 core-logic; 符号 `BatchedMinNewTokensPenalizer._prepare`): 核心修复: 合并模型配置 EOS token IDs 到停止 token 集合, 并过滤 `None` 值避免 tensor 转换崩溃。
- `test/registered/unit/sampling/test_penaltylib.py` (模块 采样测试; 类别 test; 类型 test-coverage; 符号 `test_blocks_model_eos_without_tokenizer_eos`): 新增测试覆盖

tokenizer 无 EOS 但模型配置有 EOS 的场景，保证修复有效。

关键符号: BatchedMinNewTokensPenalizer._prepare,
test_blocks_model_eos_without_tokenizer_eos

关键源码片段

python/sglang/srt/sampling/penaltylib/min_new_tokens.py

核心修复: 合并模型配置 EOS token IDs 到停止 token 集合, 并过滤 None 值避免 tensor 转换崩溃。

```
# python/sglang/srt/sampling/penaltylib/min_new_tokens.py
class BatchedMinNewTokensPenalizer(_BatchedPenalizer):
    def _prepare(self):
        # ... 省略 min_new_tokens 构造 ...

        padded_stop_token_ids = torch.nn.utils.rnn.pad_sequence(
            sequences=[
                torch.tensor(
                    data=[
                        token_id
                        for token_id in (
                            (req.sampling_params.stop_token_ids or set())
                            | (req.eos_token_ids or set()) # 模型配置的 EOS token ID
                            | (req.tokenizer.additional_stop_token_ids or set())
                            | (
                                {req.tokenizer.eos_token_id}
                                if req.tokenizer.eos_token_id is not None
                                else set()
                            )
                        )
                    ],
                    if token_id is not None # 过滤 None 避免 tensor 转换崩溃
                ),
                dtype=torch.int64,
                device=self.orchestrator.device,
            )
            for req in self.orchestrator.reqs()
        ],
        batch_first=True,
        padding_value=self.orchestrator.vocab_size,
    )

    # 后续处理 stop_token_penalties 和 len_output_tokens ...
```

test/registered/unit/sampling/test_penaltylib.py

新增测试覆盖 tokenizer 无 EOS 但模型配置有 EOS 的场景，保证修复有效。

```
# test/registered/unit/sampling/test_penaltylib.py
class TestBatchedMinNewTokensPenalizer(CustomTestCase):
```

```
def test_blocks_model_eos_without_tokenizer_eos(self):
    """Model-config EOS remains blocked when tokenizer EOS metadata is missing."""
    req = _make_req(min_tokens=3, eos_id=None) # tokenizer EOS 缺失
    req.eos_token_ids = {6} # 模型配置的 EOS
    batch = _make_batch([req])
    orch = BatchedPenalizerOrchestrator(
        VOCAB_SIZE, batch, {BatchedMinNewTokensPenalizer}
    )
    pen = orch.penalizers[BatchedMinNewTokensPenalizer]

    logits = torch.zeros(1, VOCAB_SIZE)
    pen.apply(logits)
    # 模型 EOS (6) 应被抑制为 -inf
    self.assertTrue(torch.isneginf(logits[0, 6]))
```

评论区精华

Review 中无评论，主要讨论发生在 PR 评论区：维护者 [nvpohanh](#) 要求作者 [milesial](#) 多次 rebase/merge 到最新 main 分支，作者确认无冲突。最终 [ch-wan](#) 批准合并。无设计或正确性争议。

- rebasing 要求 (other): 作者完成 rebase，无冲突。

风险与影响

- 风险：风险较低。主要风险是引入 req.eos_token_ids 属性依赖，如果该属性在某些路径下不存在会导致 AttributeError。但该属性源自模型配置，通常在请求对象中始终存在。过滤 None 的逻辑确保不会出现类型错误。测试覆盖了关键场景，但缺少对 req.eos_token_ids 为 None 且 tokenizer EOS 存在的回归测试。
- 影响：影响范围限定在 min-new-token 惩罚器逻辑，影响所有使用 min_new_tokens 采样参数的请求，特别是那些模型配置了 EOS 但 tokenizer 未暴露的场景。修复后避免崩溃并保持行为一致，对用户影响积极。团队维护成本低，无 API 变更。
- 风险标记：依赖新增属性，测试覆盖有限

关联脉络

- 暂无明显关联 PR