PR #27187 完整报告

sgl-project/sglang

Revert "Fix TokenizerManager crash on top_logprobs with tensor values"

合并时间：2026-06-04 05:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27187

执行摘要

- 一句话：回退 top_logprobs 张量值修复，回归旧 bug
- 推荐动作：不建议直接合入：除非有明确理由（如原修复引入了更严重的 bug），否则应暂缓。建议在 revert 后立即跟进新的修复方案，并恢复测试覆盖。若确需 revert，应在 PR 描述中详细说明原因。

功能与动机

PR body 明确说明这是对 #26825 的 revert，未提供额外动机。疑似为恢复某种行为或解决合并冲突，但无详细解释。

实现拆解

1. 恢复条件判断：在 python/sglang/srt/managers/tokenizer_manager.py 中，将 detokenize_top_logprobs_tokens 的第 2191 行从 if token_logprobs_val[i] is not None: 改回 if token_logprobs_val[i]:，重新引入对张量值的布尔求值。
2. 删除测试文件：完全移除 test/registered/unit/managers/test_tokenizer_manager_top_logprobs_tensor.py（82 行），该测试覆盖了多元素张量不崩溃、None 位置保留、普通列表路径等场景。
3. 单次提交：整个 revert 由单个提交完成，无后续补充修复。

关键文件：

- python/sglang/srt/managers/tokenizer_manager.py（模块 分词器管理；类别 source；类型 core-logic）：核心逻辑变更：恢复有 bug 的条件判断，使多元素张量在布尔求值时崩溃。
- test/registered/unit/managers/test_tokenizer_manager_top_logprobs_tensor.py（模块 测试文件；类别 test；类型 deletion；符号 _make_tokenizer_manager, TestDetokenizeTopLogprobsTensor, test_multi_element_tensor_value_does_not_crash, test_none_position_yields_none）：完全删除测试覆盖，移除对多元素张量、None 位置和普通列表路径的回归测试。

关键符号：detokenize_top_logprobs_tokens

关键源码片段

python/sglang/srt/managers/tokenizer_manager.py

核心逻辑变更：恢复有 bug 的条件判断，使多元素张量在布尔求值时崩溃。

```
# python/sglang/srt/managers/tokenizer_manager.py 第 2181-2199 行
# 注意：此处使用 if token_logprobs_val[i]: 进行布尔求值，
# 当 token_logprobs_val[i] 为多元素 torch.Tensor 时将引发 RuntimeError

def detokenize_top_logprobs_tokens(
    self,
    token_logprobs_val: List[float],
    token_logprobs_idx: List[int],
    decode_to_text: bool,
):
    ret = []
    for i in range(len(token_logprobs_val)):
        # WARNING: 该条件判断在多元素 Tensor 下会崩溃
        if token_logprobs_val[i]: # revert: 从 if ... is not None 改回
            ret.append(
                self.detokenize_logprob_tokens(
                    token_logprobs_val[i], token_logprobs_idx[i], decode_to_text
                )
            )
        else:
            ret.append(None)
    return ret
```

test/registered/unit/managers/test_tokenizer_manager_top_logprobs_tensor.py

完全删除测试覆盖，移除对多元素张量、None 位置和普通列表路径的回归测试。

```
# 已删除的文件：test/registered/unit/managers/test_tokenizer_manager_top_logprobs_tensor.py
# 包含以下关键测试：

import unittest
import torch
from sglang.test.ci.ci_register import register_cpu_ci
from sglang.test.test_utils import CustomTestCase
from sglang.srt.managers.tokenizer_manager import TokenizerManager

register_cpu_ci(est_time=5, suite="base-a-test-cpu")

def _make_tokenizer_manager() -> TokenizerManager:
    # 绕过 __init__ 创建 TokenizerManager 实例
    return TokenizerManager.__new__(TokenizerManager)

class TestDetokenizeTopLogprobsTensor(CustomTestCase):
    def test_multi_element_tensor_value_does_not_crash(self):
        """多元素 Tensor 不应导致崩溃"""
        tm = _make_tokenizer_manager()
```

```
        val = [torch.tensor([-0.1, -0.2, -0.3])]
        idx = [[10, 20, 30]]
        ret = tm.detokenize_top_logprobs_tokens(val, idx, decode_to_text=False)
        self.assertEqual(len(ret), 1)
        self.assertEqual(ret[0], [(-0.1, 10, None), (-0.2, 20, None), (-0.3, 30, None)])

    def test_none_position_yields_none(self):
        """None 标记位应保持 None"""
        tm = _make_tokenizer_manager()
        val = [None, torch.tensor([-0.5, -0.6])]
        idx = [None, [1, 2]]
        ret = tm.detokenize_top_logprobs_tokens(val, idx, decode_to_text=False)
        self.assertIsNone(ret[0])

    def test_plain_list_values_still_work(self):
        """普通 List[float] 路径不受影响"""
        tm = _make_tokenizer_manager()
        val = [[-0.1, -0.2]]
        idx = [[7, 8]]
        ret = tm.detokenize_top_logprobs_tokens(val, idx, decode_to_text=False)
        self.assertEqual(ret, [[(-0.1, 7, None), (-0.2, 8, None)]])
```

评论区精华

review 中无实质性讨论。chatgpt-codex-connector[bot] 自动评论指出：当 overlap scheduling 将 top-logprobs 作为 CPU 张量保留时（未转换为列表），请求带有 return_logprob 和 top_logprobs_num > 1 会导致 token_logprobs_val[i] 为多元素张量，布尔转换引发崩溃，建议避免对张量进行 truth-testing。

- 避免对 top-logprob 张量进行布尔求值 (correctness): 无 resolved 讨论，但该评论指出了 revert 引入的 bug。

风险与影响

- 风险：回归风险（高）：直接回退已合并且经过测试的 bugfix，在 PD 分离场景下客户端发送 top_logprobs > 0 请求时，prefill 进程可能被 SIGKILL。缺少测试覆盖：删除了专门的回归测试，后续无 CI 防护。影响面：所有使用 top_logprobs 且后端返回张量值的部署。
- 影响：用户：可能遭遇服务不可用（prefill 进程反复重启）。系统：PD 分离部署中 prefill 实例可能进入重启风暴，影响稳定性。团队：需重新审视 revert 动机，若为临时回退应补充计划。
- 风险标记：回归风险，删除测试覆盖，核心路径变更，缺少说明

关联脉络

- PR #26825 Fix TokenizerManager crash on top_logprobs with tensor values: 本 PR 直接回退该修复，重新引入原 bug。