PR #26299 完整报告

sgl-project/sglang

[PD] Fix top logprobs crash in prefill path

合并时间：2026-05-26 22:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26299

执行摘要

- 一句话：修复 PD 模式下 top_logprobs 崩溃问题
- 推荐动作：值得合入，修复逻辑清晰，测试覆盖充分。开发者在后续涉及 prefill 结果处理时应注意与 batch_result_processor 中的转换逻辑保持对称。

功能与动机

Issue #26286 报告了 PD 模式下带有 top_logprobs 的异步 prefill 请求会导致 TokenizerManager 子进程崩溃，错误信息为 'Boolean value of Tensor with more than one element is ambiguous'。原因是 prefill 路径遗漏了对几个 top_logprobs 相关字段的 .tolist() 转换，而正常 prefill/decode 路径中已经在 batch_result_processor 中完成了转换。

实现拆解

1. 修复核心缺失转换：在 python/sglang/srt/disaggregation/prefill.py 的 process_batch_result_disagg_prefill 函数中，在原有的 next_token_logprobs 和 input_token_logprobs 转换之后，增加了对 next_token_top_logprobs_val、next_token_top_logprobs_idx 和 next_token_token_ids_logprobs_val 的逐元素 .tolist() 转换逻辑。
2. 添加端到端测试：在 test/registered/disaggregation/test_disaggregation_basic.py 中新增 test_chat_completion_top_logprobs 测试用例，通过 OpenAI 客户端发起带 logprobs=True 和 top_logprobs=5 的聊天补全请求，验证响应中 top_logprobs 字段存在、类型正确且包含有效 token。
3. 合并与 CI 验证：经过多次合并 main 分支并触发 CI，测试通过，修复生效。

关键文件：

- python/sglang/srt/disaggregation/prefill.py（模块 调度器；类别 source；类型 core-logic）：修复核心：在 prefill 结果处理函数中补充了 top_logprobs 相关字段的 .tolist() 转换，避免 Tensor 在多元素时被用作布尔值导致的崩溃。
- test/registered/disaggregation/test_disaggregation_basic.py（模块 测试；类别 test；类型 test-coverage；符号 test_chat_completion_top_logprobs）：新增端到端测试用例，验证 PD 模式下带 top_logprobs 的 chat completion 请求不会崩溃，并检查 top_logprobs 数据的正确性。

关键符号：test_chat_completion_top_logprobs

关键源码片段

python/sglang/srt/disaggregation/prefill.py

修复核心：在 prefill 结果处理函数中补充了 top_logprobs 相关字段的 .tolist() 转换，避免 Tensor 在多元素时被用作布尔值导致的崩溃。

```
# python/sglang/srt/disaggregation/prefill.py (head)

# 原有的转换逻辑保持不变
if logits_output.next_token_logprobs is not None:
    logits_output.next_token_logprobs = logits_output.next_token_logprobs.tolist()
if logits_output.input_token_logprobs is not None:
    logits_output.input_token_logprobs = tuple(
        logits_output.input_token_logprobs.tolist())

# 新增的 top_logprobs 转换，与 batch_result_processor 保持对称
if logits_output.next_token_top_logprobs_val:
    # 逐元素 tolist：每个元素的 topk 值转换为列表
    logits_output.next_token_top_logprobs_val = [
        v.tolist() for v in logits_output.next_token_top_logprobs_val
    ]
    logits_output.next_token_top_logprobs_idx = [
        x.tolist() for x in logits_output.next_token_top_logprobs_idx
    ]
if logits_output.next_token_token_ids_logprobs_val:
    logits_output.next_token_token_ids_logprobs_val = [
        v.tolist() for v in logits_output.next_token_token_ids_logprobs_val
    ]
```

test/registered/disaggregation/test_disaggregation_basic.py

新增端到端测试用例，验证 PD 模式下带 top_logprobs 的 chat completion 请求不会崩溃，并检查 top_logprobs 数据的正确性。

```
# test/registered/disaggregation/test_disaggregation_basic.py (head)

def test_chat_completion_top_logprobs(self):
    # 使用 OpenAI 客户端模拟用户请求，开启 logprobs 和 top_logprobs=5
    client = openai.Client(api_key="empty", base_url=f"{self.lb_url}/v1")
    response = client.chat.completions.create(
        model="dummy",
        messages=[
            {"role": "system", "content": "You are a helpful AI assistant."},
            {"role": "user", "content": "What is the capital of France?"},
        ],
        temperature=0,
        max_tokens=8,
        logprobs=True,
        top_logprobs=5,
    )
```

```
    # 验证响应中包含 logprobs 信息
    self.assertIsNotNone(response.choices[0].logprobs)
    content_logprobs = response.choices[0].logprobs.content
    self.assertGreater(len(content_logprobs), 0)

    # 获取第一个包含 top_logprobs 的条目，并验证其格式
    first_top_logprobs = next(
        (item.top_logprobs for item in content_logprobs if item.top_logprobs),
        None,
    )
    self.assertIsNotNone(first_top_logprobs)
    self.assertGreater(len(first_top_logprobs), 0)
    self.assertIsInstance(first_top_logprobs[0].token, str)
```

评论区精华

审核人 ShangmingCai 对修复表示认可（“Looks good.”），并主动触发 CI 运行。无其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅新增了三个字段的类型转换，与现有逻辑正交，不改变原有的控制流。但需注意，在新增的测试中使用了 dummy 模型，可能无法覆盖所有模型的 top_logprobs 行为，建议在真实模型上做额外验证。
- 影响：影响范围明确：仅修复 PD 模式下 top_logprobs 功能，不影响其他路径或功能。用户使用 PD 部署并开启 logprobs + top_logprobs 时将不再遇到 500 崩溃。
- 风险标记：核心路径变更

关联脉络

- PR #26286 [Bug] PD chat/completions with top_logprobs can crash TokenizerManager on tensor truth-value check: 关联 Issue，描述了该 bug 的触发条件和错误栈，是本 PR 修复的直接动机。