

PR #32410 完整报告

sgl-project/sglang

Fix flaky test_sampling_mask: mask length can legitimately be top_k + 1

合并时间: 2026-07-26 07:05

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32410>

执行摘要

- 一句话: 修复不稳定的 testing_mask 测试断言
- 推荐动作: 值得合并: 修复了不稳定测试, 无需 runtime 变更。

功能与动机

PR body 指出测试 `test_generate_returns_sampling_mask` 不稳定, 最近 CI 运行出现以下错误: `AssertionError: 11 not less than or equal to 10`。根因是

`Sampler._attach_sampling_mask_to_output` 有意在 mask 后追加实际采样的 token (当采样 kernel 选中刚好超出 mask 的 `torch.topk` 重建范围的 token 时), 导致合法长度为 `top_k + 1`。该测试自引入以来一直存在潜在不稳定性。

实现拆解

该PR仅修改测试文件`test/registered/sampling/test_sampling_mask.py`, 对核心逻辑无影响。
变更内容为:

- 对于 top_p 采样情况: 将 `assertLessEqual(len(sampling_mask), _TOP_K)` 改为 `assertLessEqual(len(sampling_mask), _TOP_K + 1)`。
- 对于仅 top_k 以及 top_k + top_p=1.0 的情况: 将 `assertEqual(len(sampling_mask), _TOP_K)` 改为 `assertIn(len(sampling_mask), (_TOP_K, _TOP_K + 1))`, 以容忍 +1 情况的同时保留下限保证。
- 添加注释说明 +1 的原因 (浮点累积和偏差)。

关键文件:

- `test/registered/sampling/test_sampling_mask.py` (模块 采样测试; 类别 test; 类型 test-coverage; 符号 `test_generate_returns_sampling_mask`): 唯一修改的文件: 将测试断言从严格的 `assertEqual/assertLessEqual` 放宽为允许 mask 长度为 `top_k + 1`, 以匹配 `Sampler._attach_sampling_mask_to_output` 的实际行为。

关键符号: `test_generate_returns_sampling_mask`

关键源码片段

`test/registered/sampling/test_sampling_mask.py`

唯一修改的文件：将测试断言从严格的 `assertEqual/assertLessEqual` 放宽为允许 mask 长度为 `top_k + 1`，以匹配 `Sampler._attach_sampling_mask_to_output` 的实际行为。

```
def test_generate_returns_sampling_mask(self):
    # ... setup ...
    # The mask keeps at most top_k tokens, plus possibly the actually
    # sampled token when the sampling kernel picks one just outside the
    # mask's topk reconstruction (fp cumsum divergence); see
    # Sampler._attach_sampling_mask_to_output.
    for sampling_mask in top_p_sampling_masks:
        self.assertLessEqual(len(sampling_mask), _TOP_K + 1)

    # ... other cases ...
    for sampling_mask in top_k_sampling_masks:
        self.assertIn(len(sampling_mask), (_TOP_K, _TOP_K + 1))

    for sampling_mask in top_k_top_p_one_sampling_masks:
        self.assertIn(len(sampling_mask), (_TOP_K, _TOP_K + 1))
```

评论区精华

该 PR 无 review 评论。作者 merrymercy 自行审查并批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试断言，不改变任何运行时行为。测试仍然验证 sampled token 在 mask 内 (`assertIn(output_id, sampling_mask)`)。
- 影响：影响范围极小：仅影响一个测试文件中三个断言行。消除不确定性，使 CI 更可靠。
- 风险标记：测试覆盖调整

关联脉络

- PR #27408 Introduce sampling mask feature: 该 PR 引入了 sampling mask 功能，其中的测试逻辑与本次修复相关。