PR #26534 完整报告

sgl-project/sglang

fix: use req.req_pool_idx instead of loop variable for req_to_token i···

合并时间：2026-05-29 17:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26534

执行摘要

- 一句话：修复 bench_one_batch.py 中 req_to_token 索引错用循环变量
- 推荐动作：该 PR 是一个简单但重要的正确性修复。阅读源码片段可了解 req_to_token_pool 的正确索引方式，这对 SGLang 的请求池管理工作有参考价值。

功能与动机

在 bench_one_batch.py 的 prepare_extend_inputs_for_correctness_test 中，循环变量 i 被错误地用于索引 req_to_token_pool。当请求在池中不按连续顺序存储时，会导致静默的索引不匹配——错误的 prefix KV 缓存被分配给请求，从而产生不正确的正确性测试结果。

实现拆解

1. 在 python/sglang/bench_one_batch.py 文件中的 prepare_extend_inputs_for_correctness_test 函数（第 384-393 行），将循环变量 i 替换为 req.req_pool_idx，用于从 model_runner.req_to_token_pool.req_to_token 中索引 prefix indices。
2. 其他部分不变。该函数用于正确性测试，从预先准备好的 request 列表中提取 extend inputs，并设置 prefix indices 以实现 KV cache 重用。

关键文件：

- python/sglang/bench_one_batch.py（模块 基准测试；类别 source；类型 core-logic；符号 prepare_extend_inputs_for_correctness_test）：唯一变更文件，修复了 prepare_extend_inputs_for_correctness_test 函数中 req_to_token 索引使用循环变量 i 而非 req.req_pool_idx 的 bug。

关键符号：prepare_extend_inputs_for_correctness_test

关键源码片段

python/sglang/bench_one_batch.py

唯一变更文件，修复了 prepare_extend_inputs_for_correctness_test 函数中 req_to_token 索引使用循环变量 i 而非 req.req_pool_idx 的 bug。

```
# python/sglang/bench_one_batch.py
# 修复前：使用循环变量 i 索引 req_to_token_pool，当请求在池中不连续存储时取错 prefix
```

```
# 修复后：使用 req.req_pool_idx 确保每个请求取回自己的 prefix indices

def prepare_extend_inputs_for_correctness_test(
    bench_args, input_ids, reqs, model_runner
):
    for i in range(len(reqs)):
        req: Req = reqs[i]
        req.fill_ids += input_ids[i][bench_args.cut_len:]
        if model_runner is not None:
            # Bug: 原本使用 i 索引，当请求不连续存储时取错 prefix
            # Fix: 使用 req.req_pool_idx 确保取回正确请求的 prefix
            req.prefix_indices = model_runner.req_to_token_pool.req_to_token[
                req.req_pool_idx, : bench_args.cut_len
            ].to(req.prefix_indices.dtype)
            req.logprob_start_len = -1
            req.set_extend_input_len(len(req.fill_ids) - len(req.prefix_indices))
    return reqs
```

评论区精华

维护者 hnyls2002 指出该脚本不被 CI 覆盖，且已不活跃（'This is not covered by CI, and actually, this script is not active'）。尽管如此，他仍感谢贡献并合并了 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。只修改了一行，且变更逻辑正确：使用 req.req_pool_idx 而不是循环索引 i，符合 SGLang 中 req_to_token_pool 的预期用法。该脚本主要用于内部基准测试，不影响生产路径。
- 影响：影响范围局限于 bench_one_batch.py 脚本的正确性测试模式。使用 --correct 标志运行基准测试的用户将获得正确的 prefix KV cache 索引，从而得到准确的输出对比。由于该脚本不被 CI 覆盖且已不活跃，实际用户影响有限。
- 风险标记：非活跃脚本，无 CI 覆盖

关联脉络

- 暂无明显关联 PR