PR #26481 完整报告

sgl-project/sglang

Fixed incorrect indexing for slot 0 compatibility

合并时间：2026-06-01 10:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26481

执行摘要

- 一句话：修复 bench_one_batch.py slot 0 索引错位
- 推荐动作：建议快速合并，修复明确且验证充分（PR body 附有测试命令）。

功能与动机

PR#24243 将 slot 0 预留为 padding 后，bench_one_batch.py 的测试工具使用循环索引 i（0, 1, 2, ...）访问 req_to_token，但实际分配的池索引从 1 开始，导致前缀索引查找错位。

实现拆解

1. 修正 fill_ids 追加方式：在第 386 行将 req.fill_ids += input_ids[i][bench_args.cut_len :] 改为 req.fill_ids.extend(...)，避免 TypeError（列表不能直接 += 一个列表片段？实际上是 Python 语义正确但可能导致预期外的行为，改用 extend 更明确）。
2. 修正 req_to_token 索引：在第 389 行将 i 替换为 req.req_pool_idx，以正确反映 slot 0 预留后的实际池索引。
3. 仅修改 python/sglang/bench_one_batch.py 一个文件，变更量小（+3/-2）。

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）：唯一修改的文件，核心逻辑修复：使用 req.req_pool_idx 替代循环索引 i，兼容 slot 0 padding。

关键符号：prepare_extend_inputs_for_correctness_test

关键源码片段

python/sglang/bench_one_batch.py

唯一修改的文件，核心逻辑修复：使用 req.req_pool_idx 替代循环索引 i，兼容 slot 0 padding。

```
# python/sglang/bench_one_batch.py: 修复后的 prepare_extend_inputs_for_correctness_test

def prepare_extend_inputs_for_correctness_test(
    bench_args, input_ids, reqs, model_runner
):
```

```
    for i in range(len(reqs)):
        req: Req = reqs[i]
        # 使用 .extend() 而非 += 以避免类型错误（列表扩展）
        req.fill_ids.extend(input_ids[i][bench_args.cut_len :])
        if model_runner is not None:
            # 使用 req.req_pool_idx 而非 i 来正确处理 slot 0 填充
            # PR#24243 将 slot 0 预留为 padding，因此实际池索引从 1 开始
            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

review 由 gemini-code-assist 自动审核确认改动正确，无额外讨论；mingfeima 和 polisettyvarma 分别批准，过程中无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更范围仅限于 bench_one_batch.py 的 prepare_extend_inputs_for_correctness_test 函数，且仅修复索引和类型问题，不影响其他功能路径。但缺少直接针对此修复的单元测试。
- 影响：影响范围限于使用 bench_one_batch.py 进行正确性测试的开发者，修正后前置索引正确，测试结果可靠。对 Intel XPU 和 CUDA 的基准测试均有正面影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #24243 Reserve slot 0 as padding in all req pools: 本 PR 修复了 PR#24243 引入 slot 0 预留后导致的索引错位问题。