

PR #27952 完整报告

sgl-project/sglang

Bump spec logprob match delta for the bf16 eagle fixture

合并时间: 2026-06-12 05:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27952>

执行摘要

- 一句话: 调高推测解码 logprob 匹配阈值以适应 bf16
- 推荐动作: 此 PR 是常规测试维护, 不值得精读, 但可关注其阈值设定逻辑作为测试健壮性设计的参考。

功能与动机

PR body 指出, 0.255 的阈值是为 fp16 校准的; fixture 迁移至 bf16 后 (#27883), decode-vs-prefill 舍入噪声本地达到 ~0.25, CI 上高达 0.36, 导致 `test_logprob_spec_v2_match` 在定时运行中间歇性失败。

实现拆解

1. 在 SpecLogprobKit 类定义中新增类属性 `logprob_match_delta = 0.5`, 并添加注释说明 bf16 噪声特性及与 fp16 的差异。
2. 将 `test_logprob_match` 和 `test_logprob_spec_v2_match` 中所有的 `self.assertLess(max_diff, 0.255)` 替换为 `self.assertLess(max_diff, self.logprob_match_delta)`。
3. 将 `test_logprob_spec_v2_match` 中 `assertAlmostEqual` 的 `delta=0.255` 替换为 `delta=self.logprob_match_delta`。

关键文件:

- `python/sglang/test/kits/spec_server_kits.py` (模块 测试工具; 类别 test; 类型 test-coverage): 这是唯一修改的文件, 将硬编码阈值提升为类属性并增大到 0.5, 以适配 bf16 精度。

关键符号: 未识别

关键源码片段

`python/sglang/test/kits/spec_server_kits.py`

这是唯一修改的文件, 将硬编码阈值提升为类属性并增大到 0.5, 以适配 bf16 精度。

```
# python/sglang/test/kits/spec_server_kits.py
class SpecLogprobKit:
    """Logprob correctness: start_len, prefill-rescore match, mixed sweep,
```

```
spec-v2 decode-vs-prefill match, and ragged token_ids_logprob."""

# Max ldecode-path - prefill-rescore logprob gap. The two paths run
# different kernels / batch shapes, so the gap is accumulated rounding
# noise of the fixture dtype: ~0.25 observed for bf16 (up to 0.36 on
# some CI runners), ~8x smaller for fp16 (3 extra mantissa bits).
logprob_match_delta = 0.5

def test_logprob_match(self):
    # ... (omitted)
    diff = np.abs(output_logprobs - output_logprobs_score)
    max_diff = np.max(diff)
    self.assertLess(max_diff, self.logprob_match_delta) # was 0.255

def test_logprob_spec_v2_match(self):
    # ... (omitted)
    max_diff = np.max(np.abs(decode_vals - score_vals))
    print(f"[round {round_idx}] logprob max_diff={max_diff:.6f}")
    self.assertLess(max_diff, self.logprob_match_delta) # was 0.255
    for tid in common_ids:
        self.assertAlmostEqual(
            dec_top[tid], scr_top[tid], delta=self.logprob_match_delta # was 0.255
        )
    for tid in dec_tid:
        self.assertAlmostEqual(
            dec_tid[tid], scr_tid[tid], delta=self.logprob_match_delta # was 0.255
        )
```

评论区精华

该 PR 无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：阈值提升至 0.5 可能掩盖真正的 logprob 回归问题，因为显著的有意义差异（如 kernel bug 导致的 >0.5 偏差）仍会被捕获，但较小偏差（<0.5）将不再触发告警。当前改动合理，因为 bf16 的内在精度限制了更严格阈值的可行性。
- 影响：仅影响测试框架 spec_server_kits.py 中的一个类，不涉及生产代码。修复了 CI 间歇性失败，提升了测试稳定性。
- 风险标记：测试阈值调整，潜在回归掩盖

关联脉络

- PR #27883 Fixture move to bf16 (推测题目未提供该 PR 具体信息，但 body 提及)：该 PR 将测试 fixture 从 fp16 迁移至 bf16，直接导致了本 PR 的阈值调整需求。