PR #24965 完整报告

sgl-project/sglang

[Spec] Fix ngram metric off-by-1 in `num_accepted_drafts_per_req_cpu`

合并时间：2026-05-12 03:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24965

执行摘要

- 一句话：修复 ngram 投机解码指标偏移
- 推荐动作：值得快速合入，修复明显。建议后续补充单元测试验证指标值。

功能与动机

ngram 路径将 verify_input.num_accepted_drafts（仅草案数）作为 accept_lens 传入，下游 scheduler_output_processor_mixin 期望包含 bonus token 的值并减 1，导致统计指标 spec_accepted_drafts 和 spec_acceptance_histogram 错误地少 1。PR body 明确指出该问题。

实现拆解

1. 定位文件：python/sglang/srt/speculative/ngram_worker.py 中 forward_batch_generation 方法的第 332 行。
2. 变更：将 accept_lens = verify_input.num_accepted_drafts 改为 accept_lens = verify_input.num_accepted_tokens。num_accepted_tokens 已在 ngram_info.py:456 计算为 num_accepted_drafts + 1，包含 bonus token。
3. 更新注释：从 "Store accept_lens for per-request metrics" 更新为 "Store accept_lens (with bonus) for per-request metrics; downstream subtracts 1 to recover drafts-only counts."，明确下游行为。
4. 未涉及测试、配置或部署配套更改。

关键文件：

- python/sglang/srt/speculative/ngram_worker.py（模块 投机解码；类别 source；类型 core-logic）：单文件修改，核心逻辑变更：修复指标中 accept_lens 的取值。

关键符号：ngram_worker.forward_batch_generation

关键源码片段

python/sglang/srt/speculative/ngram_worker.py

单文件修改，核心逻辑变更：修复指标中 accept_lens 的取值。

```
        if get_global_tracing_enabled():
            for idx, req in enumerate(batch.reqs):
```

```
                accepted = (
                    verify_input.num_accepted_drafts[idx].item()
                    if verify_input.num_accepted_drafts is not None
                    else 0
                )
                req.time_stats.set_spec_verify_end_time(accepted_tokens=accepted)

        # Store accept_lens (with bonus) for per-request metrics; downstream
        # subtracts 1 to recover drafts-only counts.
        accept_lens = verify_input.num_accepted_tokens
        if batch.return_logprob:
            add_output_logprobs_for_spec_v1(batch, verify_input, logits_output)
```

该代码块来自 forward_batch_generation 方法。关键变更在第 333 行：将 accept_lens 的赋值来源从num_accepted_drafts（仅草案数）改为num_accepted_tokens（含bonustoken），以匹配下游处理器的预期。

评论区精华

无 reviewer 讨论。仅 gemini-code-assist[bot] 自动评论确认变更内容，无反馈。PR 作者触发了一次 /rerun-test 以执行 test_ngram_speculative_decoding.py，测试通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更是简单的单行赋值替换，仅影响指标统计，不改变控制流、模型输出或采样结果。num_accepted_tokens 已在同一代码路径中计算，无新增依赖。但缺少新增测试来验证指标正确性，可能存在回归风险。
- 影响：影响范围窄，仅修复 ngram 投机解码场景下 num_accepted_drafts_per_req_cpu 指标的统计准确性。对用户：spec_accepted_drafts 和 spec_acceptance_histogram 输出将正确包含 bonus token。不影响其他投机方法（如 eagle2）。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR