PR #22497 完整报告

sgl-project/sglang

fix prefill tps log accuracy

合并时间：2026-04-12 14:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22497

执行摘要

- 一句话：修复预填充输入吞吐量日志计算错误，消除虚假 TPS 峰值。
- 推荐动作：该 PR 值得快速浏览，特别是对于关注监控指标准确性的工程师。关键设计决策是将预填充吞吐量计算与解码阶段逻辑对齐，体现了指标计算的一致性原则。虽然变更简单，但 PR 描述中的历史演进分析具有教育价值。

功能与动机

根据 PR 描述，预填充输入吞吐量指标经历了三个阶段：最初版本始终为零（缺少累加器递增），PR #7245 的修复引入了令牌 / 时间错位（使用上一批次的令牌数与当前批次的时间间隔），导致吞吐量读数不准确。当前 PR 旨在解决这一错位问题，使吞吐量计算与解码阶段的逻辑保持一致，消除虚假峰值。

实现拆解

修改了 python/sglang/srt/observability/scheduler_metrics_mixin.py 文件中的 report_prefill_stats 方法。关键改动包括：1) 移除 self.last_prefill_tokens 字段，该字段用于存储上一批次的令牌数；2) 将吞吐量计算改为使用当前批次的 prefill_stats.log_input_tokens 除以当前时间间隔 gap_latency；3) 在时间间隔为零时返回 0.0 以避免除零错误。

关键文件：

- python/sglang/srt/observability/scheduler_metrics_mixin.py（模块 observability）：唯一修改的文件，包含预填充统计报告逻辑，修复了吞吐量计算错误。

关键符号：report_prefill_stats

评论区精华

Review 中未出现实质性技术讨论。唯一评论来自 HaiShaw 的批准，无具体反馈。PR 描述中详细解释了问题背景、三个阶段的历史演进、错误公式如何导致峰值以及验证结果，但未在 review 评论中展开讨论。

- 预填充吞吐量计算错误修复 (correctness): 采用当前批次的令牌数除以当前时间间隔，与解码阶段逻辑保持一致。

风险与影响

- 风险：风险较低。变更仅影响日志指标计算，不涉及核心计算逻辑。主要风险包括：1) 回归风险：如果新公式在边缘情况（如极短时间间隔）下处理不当，可能导致除零错误或异常值，但已通过 `if gap_latency > 0 else 0.0` 防护；2) 兼容性：移除 `self.last_prefill_tokens` 字段可能影响依赖该字段的代码，但根据上下文该字段仅在此处使用；3) 测试覆盖：PR 描述提到已通过基准测试验证，但未明确是否有单元测试覆盖。
- 影响：影响范围有限。直接影响：修复了预填充阶段输入吞吐量日志的准确性，消除虚假峰值，使监控指标更可靠。间接影响：提升运维和调试体验，因为日志数据更准确。对系统性能、用户功能或 API 无影响。
- 风险标记：边缘情况处理，字段移除

关联脉络

- PR #7245 未提供，但根据 PR 描述为 'minor fix': PR 描述中提到该 PR 引入了令牌 / 时间错位，是当前修复的前序变更。