PR #21793 完整报告

sgl-project/sglang

Add latency and throughput metrics to run_eval

合并时间：2026-04-01 09:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21793

执行摘要

本 PR 在 sglang 仓库的 run_eval 脚本中添加了延迟和输出吞吐量指标计算，旨在为 CI 回归检查提供性能监控基础。通过修改评估采样器来跟踪完成令牌，实现简单且风险低的增强。

功能与动机

根据 PR body，主要动机是“Track completion_tokens in ChatCompletionSampler and compute output_throughput + latency in run_eval metrics. Foundation for regression CI checks.”这意味着为了支持 CI 流程中的性能回归测试，需要收集模型推理的延迟和令牌生成速率指标。

实现拆解

关键改动涉及两个文件：

- python/sglang/test/simple_eval_common.py: 为 ChatCompletionSampler 类添加 _completion_tokens 列表属性，并在 __call__ 方法中累积每个 API 响应的完成令牌。
- python/sglang/test/run_eval.py: 在 run_eval 函数中，针对单个重复和多个重复场景，分别计算延迟（总延迟或平均延迟）和输出吞吐量（总完成令牌 / 总延迟），并更新 metrics 字典。

示例代码逻辑： `if total_completion_tokens > 0 and latency > 0: metrics["output_throughput"] = total_completion_tokens / latency`

评论区精华

review 评论为空，表明变更未经讨论直接通过，可能因为改动较小且直接。

风险与影响

风险分析：代码已处理除零情况，但需要确保 _completion_tokens 在并发评估中正确累积；影响限于测试 metrics，不改变核心功能，可能影响基于这些指标的自动化测试。

关联脉络

相关 PR #21785 同样修改了 run_eval.py 和 simple_eval_common.py，扩展了评估采样器，显示该仓库近期在加强评估工具链，以支持更全面的 CI 性能监控。