PR #25854 完整报告

sgl-project/sglang

ci(sgl-router): add PR test workflow (pre-positioned for feature PR)

合并时间：2026-05-20 15:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25854

执行摘要

- 一句话：预置 sgl-router 的 PR 测试 CI 工作流
- 推荐动作：此 PR 虽仅为一个工作流文件，但其设计思路值得阅读：
  1. 提前部署 CI 基础设施，消除大型 PR 的阻塞点。
  2. pin-free 安装 vs 固定版本的权衡。
  3. 多层流水线 + SHA 校验的供应链安全实践。
  4. 使用 continue-on-error 处理 sccache 安装失败，提升 CI 鲁棒性。 对于管理大型工程项目 CI 的读者参考价值高。

功能与动机

如 PR body 所述，将 CI 工作流从 141 个 commit 的特性 PR（#25851）中拆分出来，避免“先有鸡还是先有蛋”的问题：任何工作流配管问题会阻塞整个特性 PR 的 CI 通过。提前落地工作流，让特性 PR 到达时 CI 自动就绪，且问题可以独立修复。

实现拆解

1. 添加工作流文件：在 .github/workflows/pr-test-sgl-router.yml 中新增 347 行配置，定义触发条件（路径过滤 experimental/sgl-router/**及关键脚本）和并发控制。
2. 设计三层流水线：
   - tier-1 lint：cargo check / clippy / fmt，包括 dynamo SHA 校验（确保依赖不被意外篡改）。
   - tier-2 build+test：cargo build + cargo test --release（跳过 tokenizer parity 以避免 HF 缓存问题）。
   - tier-3 集成：k8s integration（kind + Docker 镜像构建 + pytest）、e2e（自托管 2-GPU runner）、以及 placeholder 阶段。
3. 关键配置：
   - pin-free 安装：e2e 阶段通过 scripts/ci/cuda/ci_install_dependency.sh 以 editable 模式安装本地分支的 SGLang，避免固定版本导致与当前分支的 ServerArgs 字段脱节。
   - sccache + rust-cache 分级缓存：编译缓存降级容忍，rust-cache 保证离线命中。
   - finish gate 任务：汇总上游状态，确保任何阶段失败都能阻塞最终合并。
4. 与 #25851 的关系：此工作流合并后，特性 PR 到达时即自动触发 CI，无需额外配置。

关键文件：

- .github/workflows/pr-test-sgl-router.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：新增 sgl-router 的 PR 测试工作流，包含 lint、build、test、e2e 等多层级流水线，是后续功能 PR 的 CI 基础。关键设计包括路径过滤、dynamo SHA 校验、pin-free 安装和分级缓存。

关键符号：未识别

评论区精华

无 review 评论。PR body 中讨论了以下设计决策：

- 拆分 CI 与特性 PR：避免大型 PR 的 CI 配管阻塞。
- pin-free vs pinned 安装：选择 editable 安装以跟踪分支最新 ServerArgs 变更，防止 e2e 运行在过时代码上。
- SHA 校验：对 dynamo tokenizers 依赖进行 rev 固定，防止 supply-chain 漂移。 所有决策已在实现中落地，未产生分歧。
- 暂无高价值评论线程

风险与影响

- 风险：
  - 路径过滤依赖对齐风险：如果后续特性 PR 中 experimental/sgl-router 路径结构变更（如拆分模块），工作流可能遗漏部分子路径触发，需同步更新 path filter。
  - SHA 校验维护负担：DYNAMO_TOKENIZERS_EXPECTED_REV 常量要求每更新依赖就主动修改 PR，可能被忽略导致 CI 阻塞。
  - 自托管 runner 的可用性：e2e 阶段使用 2-gpu-h100 runner，若 runner 不稳定可能影响最终合并。
  - 无实际运行验证：在合并前无法验证工作流是否正常工作，依赖特性 PR 触发后的反馈。
- 影响：
  - 用户：无直接影响，不涉及用户可见变更。
  - 系统：新增 CI 流水线，仅在 experimental/sgl-router 路径变更时消耗资源，对现有 CI 无影响。
  - 团队：路由功能开发者将自动获得 CI 反馈，加速迭代。维护者需关注路径过滤和 SHA 校验与代码仓库的同步。
  - 风险标记：路径过滤需对齐，SHA 校验维护负担

关联脉络

- PR #25851 （待确定）引入 sgl-router 实验性模块：此工作流是为该大型特性 PR 的 CI 前置部署；#25851 包含 141 个 commit，将首次创建 experimental/sgl-router 目录，触发此工作流。