

PR #34195 完整报告

sgl-project/sglang

[CI] Align rerun-test environment with the test stages

合并时间: 2026-08-12 12:37

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34195>

执行摘要

- 一句话: 对齐 rerun-test 环境变量, 补全 JIT 完整网格开关
- 推荐动作: 值得快速浏览, 了解 CI 环境一致性管理实践; 新输入 full_jit_kernel_tests 的默认关闭设计值得借鉴。

功能与动机

PR body 指出 rerun-test.yml 只设置了 _pr-test-stage.yml 中 12 个环境变量中的 4 个, 导致 /rerun-test 可能通过而 CI 仍失败; 尤其 SGLANG_JIT_KERNEL_RUN_FULL_TESTS 未设置, 使 27 个夜间内核注册在缩减网格上重跑。

实现拆解

1. 在 .github/workflows/rerun-test.yml 的 on.workflow_call.inputs 中新增布尔输入 full_jit_kernel_tests, 默认 false, 用于控制是否对 jit_kernel 测试使用完整参数网格。
2. 在 env 部分补全 5 个环境变量: NCCL_NVLS_ENABLE: "0"、HF_HUB_DOWNLOAD_TIMEOUT: 300、HF_HUB_ETAG_TIMEOUT: 300、SGLANG_JIT_KERNEL_RUN_FULL_TESTS (由 inputs.full_jit_kernel_tests && '1' || '0' 动态生成)、IS_H200 (根据 inputs.runs_on == '8-gpu-h200' 生成)。
3. 刻意不设置 SGLANG_PRECISION_* 相关变量, 注释说明原因: 该测试会推送共享基线, rerun 若持有写令牌可能污染全局比较基准。
4. 提交历史显示作者曾尝试更大范围的重构 (如检测大模型启动预算、为 perf 和 tool-call 路径增加 launch_timeout), 但最终放弃, 仅保留环境镜像方案, 体现出克制的最小变更原则。

关键文件:

- .github/workflows/rerun-test.yml (模块 工作流; 类别 infra; 类型 infrastructure): 唯一修改文件, 补全环境变量对齐并新增输入

关键符号: 未识别

关键源码片段

[.github/workflows/rerun-test.yml](#)

唯一修改文件, 补全环境变量对齐并新增输入

```

# .github/workflows/rerun-test.yml (局部)
on:
  workflow_call:
    inputs:
      # 新增: 是否对 jit_kernel 测试使用完整参数网格
      full_jit_kernel_tests:
        description: "Sweep jit_kernel tests over their full parameter grids, the way scheduled
        suites do."
        required: false
        type: boolean
        default: false

env:
  SGLANG_IS_IN_CI: true
  SGLANG_ENABLE_ASYNC_ASSERT: true
  SGLANG_CUDA_COREDUMP: "1"
  SGLANG_JIT_DEEPGEMM_FAST_WARMUP: true

  # 与 _pr-test-stage.yml 对齐, 确保 rerun 环境一致
  NCCL_NVLS_ENABLE: "0"
  HF_HUB_DOWNLOAD_TIMEOUT: 300
  HF_HUB_ETAG_TIMEOUT: 300

  # 完整 JIT 网格默认关闭 (匹配每次提交阶段), 夜间调度需显式开启
  SGLANG_JIT_KERNEL_RUN_FULL_TESTS: ${ inputs.full_jit_kernel_tests && '1' || '0' }

  # 标识 H200 运行器, 供测试内部分支使用
  IS_H200: ${ inputs.runs_on == '8-gpu-h200' && '1' || '0' }

  # SGLANG_PRECISION_* 刻意不设置: 该测试会推送共享基线,
  # rerun 持有写令牌时可能污染全局比较基线

```

评论区精华

本 PR 没有正式的 review 评论, 但提交历史展示了设计收敛过程: 从最初的对齐环境, 到尝试检测大模型启动预算、增加超时支持, 再到最终只保留环境镜像。这种“最小必要变更”的做法值得借鉴, 避免将复杂逻辑引入 CI 脚本。

- 暂无高价值评论线程

风险与影响

- 风险: 环境变量对齐可能改变部分测试在 rerun 下的行为, 例如 NCCL_NVLS_ENABLE=0 会影响 NCCL 通信路径, 但这与 CI 阶段一致, 是预期效果。
SGLANG_JIT_KERNEL_RUN_FULL_TESTS 默认关闭, 若用户 rerun 夜间完整网格测试失败, 需要显式传 full_jit_kernel_tests=true, 否则仍会使用缩减网格, 存在误用可能。
IS_H200 仅对 H200 运行器置位, 其他运行器为空字符串, 若测试代码未防御性处理可能产生误判, 但该变量仅用于分支判断, 风险较低。未来 _pr-test-stage.yml 增加新环境变量

时，此文件可能再次漂移，需要依赖注释和审查来维持同步。

- 影响：对开发者：/rerun-test 命令更准确，减少误导；对维护者：新增一个输入参数需了解；对 CI 系统：提高复现一致性，降低调试成本。
- 风险标记：CI 环境行为变化，新输入需显式开启，环境变量可能再次漂移

关联脉络

- PR #34472 [Misc] Sanitize the structure of environ.py: 同为环境变量治理，一个在代码层，一个在 CI 层，方向互补。