# PR #25322 完整报告

sgl-project/sglang

Deprecate /rerun-stage; scrub CUDA target_stage infra

合并时间：2026-05-15 17:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25322

## 废弃 /rerun-stage 并清理 CUDA CI 基础设施

### 执行摘要

该 PR 废弃了低使用率的 /rerun-stage 斜杠命令，并从 CUDA 工作流链中系统性地移除了所有相关的 target_stage、pr_head_sha 和 include_wheel_build 基础设施。核心变更包括删除 Python 命令处理器中的 handle_rerun_stage() 函数（约 213 行），以及简化 8 个 GitHub Actions 工作流，总计减少 604 行代码。AMD 工作流保持不变。该变更基于使用数据，引导用户使用更精确的 /rerun-test 和 /rerun-failed-ci 命令。

### 功能与动机

/rerun-stage 命令允许 CI 重新运行特定阶段，但其粒度太粗，无法映射到单个功能。根据过去六个月的统计，该命令仅被使用 ~111 次（主要来自两位贡献者），而 /rerun-test 有 268 次，/rerun-failed-ci 有 320 次。因此，团队决定废弃它以简化 CI 工作流，并鼓励使用更精确的重新运行方式。

### 实现拆解

1. 废弃命令处理器（scripts/ci/utils/slash_command_handler.py）：删除了 handle_rerun_stage() 函数，该函数负责验证阶段名称、权限检查并触发 workflow_dispatch。现在，当收到 /rerun-stage 时，会通过 -1 反应和一条注释告知用户该命令已废弃，并指导他们使用替代命令。AMD 用户会被特别引导至 Actions UI 下拉菜单。
2. 清理 CUDA 工作流链：
   - 删除了所有工作流中 on.workflow_dispatch.inputs 下的 target_stage、pr_head_sha 和 include_wheel_build。
   - 移除了动态 run-name（之前用于标识 /rerun-stage 的运行）。
   - 简化了并发组，不再包括 pr_head_sha 和 target_stage。
   - 在 _pr-test-check-changes.yml 中，删除了 API 驱动的变化检测步骤（filter-api）和 validate-target-stage 步骤，简化为仅使用 paths-filter。sgl_kernel_raw 输出被移除。
   - 在所有工作流中，将 checkout ref 从 ${{ inputs.pr_head_sha || inputs.git_ref || github.sha }} 简化为 ${{ inputs.git_ref || github.sha }}。
   - 在 pr-test-multimodal-gen.yml 中，移除了所有 inputs.target_stage == '...' 条件分支，作业触发现在仅依赖于 inputs.multimodal_gen。

3. 验证 AMD 路径不变：AMD 工作流使用独立的 target_stage 输入，并通过 UI 下拉菜单触发，因此未被修改。

无（清理工作流不涉及复杂的代码逻辑）。

## 评论区精华

该 PR 没有收到人类审查评论。作者通过触发 /rerun-stage 命令验证了废弃消息的正常工作。

## 风险与影响

- 风险：依赖 /rerun-stage 的自动化脚本将不再有效，废弃消息会阻止静默失败。AMD 用户不受影响。CUDA 工作流简化后，未来若需恢复该功能需重新实现。
- 影响：开发者需要适应新的重新运行方式；CI 维护者受益于更简单的代码。整体影响中低。

## 关联脉络

该 PR 与 #25320（CI 调度改进）属于同一系列 CI 重构工作，共同推动 CI 工作流的现代化和简化。此外，它与 #25252（类型修复）和 #25316（测试文件移动）等清理工作一同体现了对代码健康度的关注。