

PR #2205 完整报告

radixark/miles

ci: authenticate CPU Hugging Face downloads

合并时间: 2026-08-05 11:49

原文链接: <http://prhub.com.cn/radixark/miles/pull/2205>

执行摘要

- 一句话: CPU 测试注入 HF_TOKEN, 修复 Hub 下载 429 限流
- 推荐动作: PR 内容简单直接, 不值得深入精读; 但 secret 最小作用域注入的设计值得借鉴, 即只把认证信息给到真正需要的步骤, 避免扩散到整个 job 或 workflow。

功能与动机

PR body 指出 `stage-b-cpu / run-cpu` 在 `pytest collection` 阶段加载

`Qwen/Qwen3-0.6B/config.json` 时遭遇 HTTP 429, 根因是 `run-cpu` 未向 `Run tests` 提供认证环境, `AutoConfig.from_pretrained` 以匿名配额访问 Hugging Face Hub 而触发限流。

实现拆解

1. 定位问题: 检查 `_run-ci.yml` 中的 `run-cpu` 作业, 确认 `Run tests` 步骤 (第 271 行附近) 没有设置任何 Hub 认证环境变量, 导致匿名配额被耗尽。
2. 注入认证变量: 在该步骤新增 `env` 块, 内容为 `HF_TOKEN: ${ secrets.HF_TOKEN }`。GitHub Actions 会将该 `secret` 作为环境变量提供给 `${ inputs.execute_command }` 运行的过程。
3. 验证与审查: 作者运行 `pre-commit run --files .github/workflows/_run-ci.yml` (含 `check yml`) 通过格式校验; reviewer `yueming-yuan` 直接 approve, 无评论。未新增测试文件, 因为该变更由 CI 自身运行结果验证。

关键文件:

- `.github/workflows/_run-ci.yml` (模块 工作流; 类别 `infra`; 类型 `configuration`): 在 `run-cpu` 作业的 `Run tests` 步骤中新增 `env.HF_TOKEN`, 修复 CPU 测试收集阶段匿名访问 Hub 被限流的问题, 是整个 PR 唯一的变更文件。

关键符号: 未识别

关键源码片段

`.github/workflows/_run-ci.yml`

在 `run-cpu` 作业的 `Run tests` 步骤中新增 `env.HF_TOKEN`, 修复 CPU 测试收集阶段匿名访问 Hub 被限流的问题, 是整个 PR 唯一的变更文件。

`.github/workflows/_run-ci.yml` 中 `run-cpu` 作业的关键步骤

```
- name: Run tests
  shell: bash
  env:
    HF_TOKEN: ${ secrets.HF_TOKEN } # 将已有 secret 仅注入测试步骤，认证 Hub 请求
  run: ${ inputs.execute_command }
```

评论区精华

本次 PR 没有产生任何 review 评论，仅有 yueming-yuan 的一次 APPROVED 空评论。作者在 body 的 Review Focus 中明确请求审查 `run-cpu` 的 `Run tests` 步骤 secret 作用域，但 reviewers 未提出异议，说明该改动被认可。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，但存在两点需注意：
 - fork PR 场景：GitHub Actions 默认不向 fork 传递 secrets，secrets.HF_TOKEN 会解析为空字符串，等于保持原匿名行为，不会新增失败，但也无法修复限流问题。
 - 打印风险：安全性依赖 inputs.execute_command 不会将环境变量打印到日志；由于 token 只注入测试步骤且 GitHub 会自动遮蔽 secret，泄露风险很小。
 - 影响范围：该改动不影响其他 job，作用域严格限定在 CPU 测试步骤。
 - 影响：影响面集中在 CI 基础设施：修复 stage-b-cpu / run-cpu 的偶发 429 失败，提升所有 PR 的 CPU 测试稳定性；对用户功能和训练代码无任何影响。团队收益是减少 CI 噪音和重跑成本。
- 风险标记：fork PR 无 secrets 时保持原行为，token 注入依赖 execute_command 不打印

关联脉络

- PR #2131 ci(docker): rebuild scheduled images at least once every 24h: 同为 CI 基础设施稳定性改进，处理构建 / 调度层面的超时与可靠性问题，与本 PR 目标一致。