PR #27092 完整报告

sgl-project/sglang

ci: cache HF hub for base-a-test-cpu to avoid Hub 429 flakes

合并时间：2026-06-04 16:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27092

执行摘要

- 一句话：为 CPU CI 添加 HuggingFace 缓存
- 推荐动作：建议合入。该 PR 采用了一致且成熟的 actions/cache 模式（与仓库其他 job 类似），可显著提高 CI 稳定性。值得关注的是其 cache key 设计：使用 github.run_id 做滚动 key，每次新 run 都会创建新 cache，同时使用 restore-keys 回退到旧 cache，既保证了增量更新又提供了回退路径。

功能与动机

base-a-test-cpu 运行在临时 GitHub Hosted Runner 上，每次运行都会从 Hub 重新下载 tokenizer/model，偶尔触发 HTTP 429 Too Many Requests 导致测试失败，与 PR 代码无关。PR body 中给出了具体失败示例：httpx.HTTPStatusError: Client error '429 Too Many Requests'。引入缓存后，热运行从 GitHub Cache 恢复模型，不再依赖网络。

实现拆解

1. 设置环境变量 HF_HOME：在 pr-test.yml 的 base-a-test-cpu job 中添加 env.HF_HOME: ${{ github.workspace }}/.hf-cache，让 huggingface 库将缓存写入该目录。
2. 引入 actions/cache@v4：新增 "Cache HF hub" step，使用滚动 key hf-cpu-${{ matrix.partition }}-${{ github.run_id }} 和 restore-keys hf-cpu-${{ matrix.partition }}-。
3. 缓存路径：path: ${{ github.workspace }}/.hf-cache，与 HF_HOME 一致。
4. 无其他文件修改：仅修改 .github/workflows/pr-test.yml 一个文件，新增 12 行，无删除。
5. 配套改动：无；测试代码无需改动，tokenizer 加载自动使用 HF_HOME。

关键文件：

- .github/workflows/pr-test.yml（模块 CI 配置；类别 infra；类型 infrastructure）：唯一修改的文件，在 base-a-test-cpu job 中添加 HF 缓存逻辑，新增 12 行配置。

关键符号：未识别

关键源码片段

.github/workflows/pr-test.yml

唯一修改的文件，在 base-a-test-cpu job 中添加 HF 缓存逻辑，新增 12 行配置。

```
# .github/workflows/pr-test.yml (partial)
# 在 base-a-test-cpu job 的 env 中设置 HF_HOME
env:
  HF_HOME: ${{ github.workspace }}/.hf-cache # 指定缓存目录

# ... 安装依赖后，添加 cache step
- name: Cache HF hub
  uses: actions/cache@v4
  with:
    path: ${{ github.workspace }}/.hf-cache # 与 HF_HOME 一致
    key: hf-cpu-${{ matrix.partition }}-${{ github.run_id }} # 每次运行新建 key
    restore-keys: |
      hf-cpu-${{ matrix.partition }}-  # 前序 key 匹配到的旧缓存作为 fallback
```

评论区精华

无 review 评论（reviewer 直接 approve），仅有一条 alisonshao 发表的 issue 评论提及另一个相关的 CI 失败链接，表明该问题普遍存在。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更局限于 CI 基础设施，仅当 cache step 失败时才退化为无缓存下载（仍能正常工作）。潜在问题包括：
  - Cache key 未包含 .hf-cache 内容哈希，若模型或 tokenizer 文件在 cache 中被污染，恢复后可能使用错误文件。但 .hf-cache 由 huggingface 库管理，污染概率极低。
  - 若 cache 命中但路径权限问题导致 HF_HOME 指定的目录不可写，后续下载会报错；但 cache 恢复后目录权限通常正确。
  - 影响：
  - 用户 / 开发者：无直接用户影响；开发者本地运行 CI 时不再因 429 误报失败。
  - CI 系统：减少了 base-a-test-cpu job 的网络依赖和运行时间（热运行节省了模型下载时间）。
  - GitHub Cache 存储：每个 partition 占用约数百 MB 至数 GB 缓存，Cache 7 天内无访问则自动清理，成本极低。
  - 风险标记：基础设施变更，缓存不一致极低风险

关联脉络

- 暂无明显关联 PR