PR #1714 完整报告

THUDM/slime

fix: http_utils. disable system proxy for internal SGLang httpx clients

合并时间：2026-03-22 16:27

原文链接：http://prhub.com.cn/THUDM/slime/pull/1714

执行摘要

本 PR 修复了在集群环境中 httpx 客户端默认使用系统代理导致内部 SGLang HTTP 调用失败的 bug，通过设置 trust_env=False 确保内部通信不被误路由，提高系统可靠性，属于有意义的改进。

功能与动机

在部署了代理环境的集群中（例如用于 W&B 日志记录），httpx.AsyncClient 默认信任环境变量（trust_env=True），导致所有内部 SGLang HTTP 请求（如 /generate、/update_weights_from_distributed、/health）被错误地通过代理路由，引发 503 Service Unavailable 错误。标准解决方法（在 no_proxy 中添加 CIDR）对 httpx 无效，因为其不支持 CIDR 表示法。因此，需要禁用这些内部客户端的代理功能以解决通信故障。

实现拆解

修改仅限于 slime/utils/http_utils.py 文件，在两个位置添加 trust_env=False：

- 在 init_http_client() 函数中，初始化 _http_client 时设置。
- 在 _RayDistributedPost.__init__ 方法中，初始化 self._client 时设置。

这些客户端专门用于集群内部的 SGLang 通信，确保请求直接发送到目标服务，而不经过外部代理。

评论区精华

本次 PR 没有 review 讨论，变更由作者直接提交并合并，表明问题明确且修复方案简单直接，无需额外评审。

风险与影响

风险：

- 如果未来这些客户端需要访问外部服务，禁用代理可能导致连接问题；但根据设计，它们仅用于内部通信，风险可控。
- 未新增测试覆盖，可能未验证其他边缘场景，如代理配置变化或网络异常。

影响：

- 修复了内部 SGLang 通信错误，确保生成、权重更新和健康检查等功能正常，提升用户体验。
- 对系统性能有积极影响，避免了不必要的代理路由，减少潜在网络延迟和错误。

关联脉络

与 PR #1768（"Fix uploading sglang metrics to wandb"）相关，两者都处理集群环境下的外部服务访问（如 W&B）和代理配置问题，反映了对 httpx 库和环境变量交互的持续优化，提示团队在类似场景中需注意代理设置。