

PR #1993 完整报告

THUDM/slime

Patch sglang 0.5.12.post1 for delta sync

合并时间: 2026-05-30 13:59

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1993>

执行摘要

- 一句话: 更新 sglang patch 支持 delta 同步
- 推荐动作: 该 PR 是基础设施适配, 无需精读, 但值得关注其与 #1991 的关联, 了解 delta 权重更新的完整链路。

功能与动机

支持 delta 权重同步功能, 使 sglang 推理服务能够正确加载和验证 slime 生成的增量权重文件。

实现拆解

1. EAGLE 工作器传参: 在 sglang/srt/speculative/eagle_worker_v2.py 和 multi_layer_eagle_worker_v2.py 的 update_weights_from_disk 调用中添加 files=recv_req.files 参数, 使 delta 权重文件能够被传递到权重加载逻辑。
2. 安全反序列化放行: 在 sglang/srt/utils/common.py 的 SafeUnpickler.ALLOW_CLASSES 中添加 "slime." 前缀, 允许反序列化 slime 模块的类, 避免 unpickle 安全限制。
3. 权重检查跳过缓存: 在 sglang/srt/utils/weight_checker.py 的 _check_tensors 中跳过 cos_sin_cache 张量检查, 因为该缓存由形状和 dtype 决定, 不同实现可能形状不同。

关键文件:

- docker/patch/latest/sglang.patch (模块 docker 补丁; 类别 test; 类型 test-coverage) : 唯一变更文件, 包含三处核心修改: 传递 files 参数、放行 slime 模块、跳过 cos_sin_cache 检查。

关键符号: update_weights_from_disk, SafeUnpickler, _check_tensors

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。修改集中在 slang 补丁文件，不会影响 slime 核心逻辑。SafeUnpickler 放行 slime 模块可能引入反序列化风险，但仅限于 slime 自身模块类，且 slime 为可信来源。
- 影响：仅影响使用 slang 作为推理后端的 delta 权重同步场景。对不使用 delta 同步的用户无影响。
- 风险标记：反序列化风险

关联脉络

- PR #1991 [ci] Add e2e test for delta weight update: 为 delta 权重更新添加了端到端测试，本 PR 的 slang patch 正是为了支持 delta 同步功能而设。