

PR #6915 完整报告

verl-project/verl

[docs] add SGLang FP8 ignored layers usage

合并时间: 2026-07-03 00:09

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6915>

执行摘要

- 一句话: 新增 SGLang FP8 忽略层文档
- 推荐动作: 值得合并, 是对已有功能的必要文档补充。无需精读, 但可作为 FP8 使用参考。

功能与动机

PR body 明确指出这是对 #6906 的文档跟进, 该 PR 在 SGLang 启动路径和 actor 权重同步路径中添加了 SGLang FP8 忽略层支持, 但用户仍不清楚如何使用。本 PR 旨在填补这一文档空白。

实现拆解

1. 在 docs/low_precision/fp8.md 的 FP8 使用章节下新增 ##### Skipping layers in SGLang FP8 rollout 小节。
2. 提供两种配置方式:
 - 环境变量: 设置 SGLANG_FP8_IGNORED_LAYERS, 支持逗号分隔多个模块名。
 - 模型配置: 在模型 quantization_config 的 ignored_layers 字段中配置, 支持 re: 正则表达式。
3. 明确说明跳过层的作用: 保持原模型精度而非转为 FP8, 适用于不兼容块状 FP8 权重量化的层或希望保持更高精度的模块。
4. 无代码 / 测试变更, 纯文档新增。

关键文件:

- docs/low_precision/fp8.md (模块 低精度文档; 类别 docs; 类型 documentation): 唯一变更文件, 新增了 SGLang FP8 忽略层的完整使用文档, 包括环境变量和模型配置两种方式。

关键符号: 未识别

评论区精华

主要讨论来自 reviewer sophiayyya 的确认提问, 询问该修改是否与 PR #6906 相关。作者 gem-mint 确认是的, 并解释了这是 #6906 的文档跟进, 说明用户可以通过 `SGLANG_FP8_IGNORED_LAYERS` 或 `quantization_config` 传递规则。讨论简洁明确, 无争议。

- 文档与 #6906 的关系确认 (documentation): 明确是 #6906 的文档补充, 内容一致。

风险与影响

- 风险: 纯文档变更, 无任何代码或配置修改, 风险极低。仅需确保文档中描述的用法与 #6906 中的实际实现一致。
- 影响: 对用户: 提供了清晰的 FP8 忽略层配置指南, 降低了功能使用门槛。对系统: 无运行态影响。对团队: 维护成本极低。影响范围限于文档读者。
- 风险标记: 纯文档变更

关联脉络

- PR #6906 [rollout] fix: support SGLang FP8 ignored layers for Qwen3.x
GatedDeltaNet in rollout: 本 PR 是 #6906 的文档跟进。#6906 实现了 SGLang FP8 忽略层的功能, 本 PR 为该功能添加使用说明。