PR #21496 完整报告

sgl-project/sglang

Revert "bugfix for weight loading for qwen3-next"

合并时间：2026-03-27 07:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21496

执行摘要

PR #21496 回退了 #21313 中对 Qwen3-next 模型权重加载属性的修改，将 _weight_loader 恢复为 weight_loader，以修正可能的错误修复。此变更仅影响特定模型初始化，建议关注测试覆盖，避免回归问题。

功能与动机

本 PR 的动机是回退之前的 bugfix (#21313)，推测其可能不正确或引起新问题。PR body 仅说明回退，无额外解释，旨在恢复原始属性名以确保权重加载正确性。

实现拆解

实现集中在 python/sglang/srt/models/qwen3_next.py 文件，具体修改如下：

- 在 __init__ 方法中，将 self.in_proj_qkvz.weight._weight_loader 改回 self.in_proj_qkvz.weight.weight_loader。
- 同样处理 self.in_proj_ba.weight 属性，从 _weight_loader 恢复为 weight_loader。

这是对 #21313 变更的完全回退，无其他调整。

评论区精华

由于没有 review 评论，无讨论内容可提炼。

风险与影响

- 风险：回退可能重新引入原始的权重加载 bug，导致模型初始化失败或权重加载不正确；变更简单但涉及核心路径，需验证兼容性。
- 影响：仅影响 Qwen3-next 模型的权重加载逻辑，影响范围有限，但若错误发生，可能影响模型功能，需通过测试确保正确性。

关联脉络

直接关联 PR #21313，该 PR 最初修复了权重加载 bug。本 PR 的回退揭示了该模型权重加载逻辑的调试迭代过程，可能反映之前修复的不稳定或验证不足，值得追踪后续相关变更。