PR #21313 完整报告

sgl-project/sglang

bugfix for weight loading for qwen3-next

合并时间：2026-03-26 21:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21313

执行摘要

该 PR 通过将 weight_loader 属性更改为 _weight_loader 来修复 Qwen3-next 模型量化权重加载错误，解决了因 @property 无 setter 导致的初始化失败，但设计上存在封装破坏风险，且后续被回退。

功能与动机

动机源于加载量化权重（如 w8a8）时，self.in_proj_qkvz.weight.weight_loader 是一个只读 @property 属性，无法设置新实现，导致错误。PR body 中提供了错误截图，显示分配属性时失败。

实现拆解

在 python/sglang/srt/models/qwen3_next.py 文件的 __init__ 方法中，修改了两行代码：

- 将 self.in_proj_qkvz.weight.weight_loader = self._make_packed_weight_loader(self.in_proj_qkvz) 改为 self.in_proj_qkvz.weight._weight_loader = ...
- 同样修改 self.in_proj_ba.weight.weight_loader 为 self.in_proj_ba.weight._weight_loader

评论区精华

gemini-code-assist[bot]评论指出：

> "Directly assigning to a 'private' attribute like _weight_loader breaks encapsulation and makes the code brittle against dependency updates..."

讨论强调直接访问私有属性是一个设计问题，建议使用公共 API，但此建议未被明确采纳。

风险与影响

- 设计风险：破坏封装，增加代码对内部实现的依赖。
- 回归风险：可能引入新 bug，Issue 评论提到可能破坏模型准确性。
- 影响范围：仅影响 Qwen3-next 模型的初始化，对使用量化权重的用户关键。

关联脉络

与此 PR 直接相关的是 PR #21496，它回退了此变更，将 _weight_loader 恢复为 weight_loader，表明原修复可能存在问题或需进一步调整。这揭示了权重加载机制的持续优化需求。