PR #21662 完整报告

sgl-project/sglang

[Fix] Fix weight_loader property assignment for qwen3-next FP8 models

合并时间：2026-03-30 16:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21662

执行摘要

本 PR 修复了 Qwen3-Next FP8 模型在加载权重时因 weight_loader 属性为只读而导致的 AttributeError，通过引入助手函数 _override_weight_loader 正确处理量化参数，使模型能正常启动，影响限于该模型类型。

功能与动机

Issue #21638 报告了在加载 Qwen/Qwen3-Coder-Next-FP8 权重时出现的 bug，错误信息为 AttributeError: property 'weight_loader' of 'ModelWeightParameter' object has no setter。原因是 BasevLLMParameter.weight_loader 是只读属性，无法直接赋值，导致量化模型加载失败。

实现拆解

修改文件 python/sglang/srt/models/qwen3_next.py，主要改动包括：

- 在 __init__ 方法中，将原本的直接赋值（如 self.in_proj_qkvz.weight.weight_loader = ...）替换为调用 _override_weight_loader 方法。
- 新增静态方法 _override_weight_loader，其核心逻辑检测参数类型：

评论区精华

由于没有正式 review 评论，Issue 评论中 ranjiewen 提出："how to deal with 'weight_scale_inv' parameter?"，该问题未在 PR 中直接解决，表明可能还有其他量化相关参数需后续处理。

风险与影响

风险：助手函数需准确区分量化与非量化参数，否则可能引发其他模型加载错误；但现有测试（如 test_qwen3_next_models.py）已覆盖，回归风险低。

影响：修复后，Qwen3-Coder-Next-FP8 模型可正常加载，提升了量化模型支持；系统其他部分不受影响，团队解决了特定部署问题。

关联脉络

与 PR #21234（支持 AMD MXFP4 Qwen3.5 模型）相关联，两者都涉及 Qwen 系列模型的量化权重加载，显示了项目在扩展量化格式支持上的持续努力。近期历史 PR 中，量化相关变更频繁，如 PR #21625 和 #18461，反映了对 FP8、MXFP 等量化技术的重视。