PR #21692 完整报告

sgl-project/sglang

[Bugfix] [NPU] Qwen3.5 with quantization fix

合并时间：2026-04-08 14:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21692

执行摘要

此 PR 修复了 NPU 平台上 Qwen3.5 量化模型因模型更新导致的映射失效问题，通过重构量化逻辑和更新模型映射确保功能正常。变更主要涉及量化层和模型定义文件，包含设计讨论和未来重构指示，属于有意义的改进。

功能与动机

PR 动机源于 Issue #21676，其中报告在更新 Qwen3.5 模型后，量化不再工作。具体来说，模型更新将 in_proj_qkv 和 in_proj_z 融合为 in_proj_qkvz，以及 in_proj_b 和 in_proj_a 融合为 in_proj_ba，导致量化映射错误。PR body 中提供了错误截图和修复后的精度测试结果（GSM8K 测试），显示修复有效。

实现拆解

实现分为三个关键部分：

1. 量化层重构（python/sglang/srt/layers/quantization/modelslim/modelslim.py）：
   - 移除 should_ignore_layer 调用，简化跳过逻辑。
   - 重构 get_linear_scheme 方法，统一参数命名为 prefix，并修复量化方案映射。
   - 代码示例：python linear_quant_schemes = [("W4A4_DYNAMIC", ModelSlimW4A4Int4), ("W8A8", ModelSlimW8A8Int8), ("W8A8_DYNAMIC", ModelSlimW8A8Int8),]
2. 模型映射更新（python/sglang/srt/models/qwen3_5.py）：
   - 扩展 packed_modules_mapping 以包括 NPU 平台（添加 _is_npu 条件），确保映射正确应用。
   - 变更示例：将 if _is_gfx95: 改为 if _is_gfx95 or _is_npu:。
3. 加载器注释添加（python/sglang/srt/model_loader/loader.py）：
   - 添加 TODO 注释，指示未来移除冗余映射代码，转向模型文件中声明的映射。

评论区精华

Review 讨论聚焦于两个核心点：

- 参数命名统一：ping1jing2 询问为何将 layer_name 改为 prefix，OrangeRedeng 解释为统一命名以匹配 get_moe_scheme，TamirBaydasov 同意此更改，指出一直使用前缀。结论

：更改被接受，提升代码一致性。

- 量化类型支持：gemini-code-assist[bot] 建议添加 W8A8 量化类型以维持向后兼容性，指出重构中丢弃了该支持。状态：建议中，PR 已合并但未直接修复，需关注潜在问题。

风险与影响

- 技术风险：量化逻辑重构可能引入回归，尤其是移除 should_ignore_layer；未支持 W8A8 量化类型可能导致现有模型失效；代码映射逻辑分散，增加维护难度。
- 影响分析：修复后，NPU 上 Qwen3.5 量化模型推理恢复正常，提升系统兼容性；团队需注意未来映射重构以避免类似 bug，代码更一致但需测试覆盖。

关联脉络

从近期历史 PR 看，此 PR 与 NPU 和量化相关功能演进相连：

- PR #21502（NPU 启用 IndexCache）涉及 NPU 后端优化，可能共享模型映射逻辑。
- PR #21240（启用 FP4 量化）涉及量化支持，技术领域相关，可对比实现方案。 整体上，仓库正持续优化 NPU 和量化功能，此 PR 是其中一环，反映代码库向更模块化和一致化方向发展。