PR #39181 完整报告

vllm-project/vllm

[Bugfix]Fix EP precision for Qwen3.5, Qwen3-Next

合并时间：2026-04-09 05:47

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39181

执行摘要

该 PR 修复了 Qwen3.5 和 Qwen3-Next 模型在启用序列并行时，共享专家权重被错误分片导致的精度问题。通过向 SharedExpert 传递 is_sequence_parallel 参数，并在序列并行时禁用张量并行，确保计算正确性。变更影响范围限于特定模型配置，风险较低，但缺乏针对性测试覆盖。

功能与动机

问题背景：当 sequence_parallel 启用时，MoE 模型中的共享专家权重会应用张量并行，但在计算结束时未执行 all-reduce 操作，导致精度损失。作者在 vllm 0.18.0 环境中使用 8 个 NVIDIA A800 GPU 测试确认此问题，并指出最新 main 分支在共享专家处理上无差异。

关键表述：> " 当 sequence_parallel 启用时，共享专家会应用张量并行且未执行 all-reduce，导致精度问题。"

实现拆解

修改涉及两个核心文件，均位于 vllm/model_executor/models/ 模块：

文件	变更点	关键代码
qwen2_moe.py	在 SharedExpert.__init__ 中添加 is_sequence_parallel 参数，并传递给 gate_up_proj 和 down_proj 的 disable_tp	```python
is_sequence_parallel: bool = False,		
...		
disable_tp=is_sequence_parallel,		
``		

文件	变更点	关键代码
qwen3_next.py	在Qwen3NextForCausalLM.__init__中将self.is_sequence_parallel传递给SharedExpert	```python
is_sequence_parallel=self.is_sequence_parallel,		
``		

逻辑说明：is_sequence_parallel 参数控制 ColumnParallelLinear 和 RowParallelLinear 的 disable_tp，当为 True 时禁用张量并行，避免权重分片，从而修复精度问题。

评论区精华

Review 讨论较少，仅 gemini-code-assist[bot] 总结变更：

> "This pull request introduces support for sequence parallelism in the Qwen2 MoE and Qwen3 Next models by adding an is_sequence_parallel parameter..."

Issue 评论中，作者提及因 DCO 问题重开 PR（关联 #38795），并请求重跑 CI 检查。vadiklyutiy 建议空提交触发 CI，最终所有检查通过后合并。无深度技术交锋。

风险与影响

技术风险：

1. 回归风险：修改了 SharedExpert 初始化签名，但参数可选，不影响现有调用；若 is_sequence_parallel 传递错误（如未正确从模型配置获取），可能导致未预期行为。
2. 测试覆盖不足：PR 未添加新测试，依赖现有测试验证，但针对序列并行 + 共享专家的特定场景可能缺乏充分测试。
3. 配置依赖：修复仅针对 Qwen3.5/Qwen3-Next，其他 MoE 模型可能存在类似问题，需额外检查。

影响评估：

- 用户影响：仅影响使用 Qwen3.5 或 Qwen3-Next 且启用序列并行的用户，修复后提升推理精度。
- 系统影响：无性能退化，通过禁用不必要分片避免计算错误。
- 团队影响：变更简单，易于维护，但需注意 MoE 模型的一致性处理。

关联脉络

- 直接关联：PR #38795 可能为同一修复的先前版本，因 DCO 问题重开。
- 功能演进：近期多个 PR 涉及 MoE 模型优化（如 #39005 重构专家目录、#39315 修复 FlashInfer MoE 崩溃），显示团队持续改进 MoE 支持。本 PR 是其中针对特定模型精度问

题的 bugfix。

- 趋势洞察：vLLM 在扩展模型支持（如 Qwen 系列）的同时，注重并行计算下的正确性，序列并行与张量并行的交互成为常见问题点。