PR #39707 完整报告

vllm-project/vllm

[Bugfix] Fix mismatch between global and local attention heads in tensor-parallel mode for param2moe model

合并时间：2026-04-14 20:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39707

PR 39707 分析报告

执行摘要

本 PR 修复了 Param2Moe 模型在张量并行模式下注意力头不匹配导致的错误计算，通过切换为本地头数和确保张量连续性，恢复了模型的正确行为，对使用该模型的用户至关重要。

功能与动机

动机是修复张量并行下 Param2moeAttention 模块的注意力计算错误。PR body 明确指出："Fix incorrect attention computation in Param2moeAttention under tensor parallelism. The Attention module was using global head counts while QKV tensors were already TP-sharded, causing a mismatch and incorrect behavior."

实现拆解

改动集中在 vllm/model_executor/models/param2moe.py 文件：

- 在 Param2MoEAttention.__init__ 中，将 Attention 模块的 num_heads 和 num_kv_heads 从全局计数改为本地计数（self.num_local_heads 和 self.num_local_kv_heads）。
- 在 Param2MoEAttention.forward 中，添加 q = q.contiguous()、k = k.contiguous()、v = v.contiguous() 确保张量连续。
- 移除冗余注释和代码格式化。

评论区精华

Review 评论较少，gemini-code-assist[bot] 简要描述了变更，DarkLight1337 直接批准，没有出现争议。讨论焦点在于代码重构的正确性。

风险与影响

风险较低，变更针对特定问题且已测试验证。但涉及核心注意力计算路径，需确保无副作用。影响范围限于 Param2Moe 模型在张量并行模式下的用户，修复后输出与 TP=1 一致。

关联脉络

从历史 PR 看，类似模型 bugfix（如 PR 39293、39688）展示了 vLLM 在支持新模型架构时的常见问题。本 PR 是 Param2Moe 模型支持线的一部分，可能关联之前引入该模型的 PR。