# PR #38517 完整报告

vllm-project/vllm

[Bugfix][Quantization] Fix PerTensorScale loading with tuple shard_id in MergedColumnParallelLinear

合并时间：2026-04-07 23:16

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38517

## 执行摘要

该 PR 修复了 MergedColumnParallelLinear.weight_loader_v2 函数中的一个关键 bug：当 PerTensorScaleParameter 遇到 tuple 类型的 shard_id（如 (0,1,2)）时，代码错误地硬编码 shard_id=0，导致量化 scale 只填充到第一个槽位，其余槽位保持垃圾值（约 -3.4e38）。这导致 Qwen3.5 模型在使用 FP8 静态逐张量量化时输出乱码。修复通过遍历 tuple 中的每个索引来正确加载 scale 值，解决了问题且无回归风险。

## 功能与动机

该 PR 旨在解决 issue #38197 中报告的问题：Qwen3.5 模型（包括 Dense 和 MoE 变体）在使用 FP8 静态逐张量量化（compressed-tensors）时输出乱码。根本原因在于 Qwen3.5 引入了 Gated Delta Net 层，其中融合投影 in_proj_qkvz 将 in_proj_qkv（3 个分区）和 in_proj_z（1 个分区）合并为一个 MergedColumnParallelLinear，其 stacked_params_mapping 定义了 tuple shard_id=(0,1,2) 表示 scale 应用于前三个分区。bug 导致只有第一个分区的 scale 被正确加载，其余分区使用垃圾 scale 值，从而破坏权重数据。

## 实现拆解

仅修改一个文件：vllm/model_executor/layers/linear.py。

关键改动：在 weight_loader_v2 函数中，当处理 PerTensorScaleParameter 且 loaded_shard_id 为 None 或 tuple 时，原代码硬编码 shard_id=0： if isinstance(param, PerTensorScaleParameter): param.load_merged_column_weight(loaded_weight=loaded_weight, shard_id=0) 修复后，新增条件判断以处理 tuple 情况： if isinstance(param, PerTensorScaleParameter): if isinstance(loaded_shard_id, tuple): for idx in loaded_shard_id: param.load_merged_column_weight( loaded_weight=loaded_weight, shard_id=idx ) else: param.load_merged_column_weight( loaded_weight=loaded_weight, shard_id=0 ) 这样，当 loaded_shard_id 为 tuple 时，会遍历每个索引调用 load_merged_column_weight，确保 scale 正确应用于所有指定分区。

## 评论区精华

review 讨论简洁但包含关键验证点：

- 审阅者 yewentao256 要求 e2e 测试："Could you also add e2e lm_eval result to make sure we don't hurt acc?"
- 作者提供量化结果：作者回应了 Qwen3.5-9B 在 GSM8K 任务上的 lm_eval 结果，显示 exact_match 从 0.876 提升到 0.884，证明修复有效且无回归。
- 最终批准：审阅者确认 "LGTM, thanks for the work!"

## 风险与影响

风险分析：

- 变更范围极小（10 行改动），逻辑清晰，直接针对已知 bug。
- 回归风险低：修复仅影响 PerTensorScaleParameter 在 tuple shard_id 场景下的行为，其他场景保持不变。
- 测试覆盖充分：作者验证了 Qwen3.5 多个模型变体（0.8B、27B、35B-A3B）和量化配置（静态 / 动态、逐张量 / 逐通道），并提供了 lm_eval 结果。
- 潜在未覆盖场景：如果其他模型使用类似 tuple shard_id 的融合投影但未被测试，可能存在影响，但鉴于变更针对性，风险较小。

影响分析：

- 用户影响：修复了 Qwen3.5 模型在 FP8 静态逐张量量化下的输出乱码问题，提升了模型可用性。
- 系统影响：仅影响权重加载逻辑，对推理性能无直接影响。
- 团队影响：解决了特定配置下的关键 bug，减少了用户支持负担。

## 关联脉络

- 直接关联 issue #38197：该 PR 直接修复此 issue，issue 提供了详细的问题描述和重现步骤。
- 与历史 PR 的关联：
  - PR #39054（修复 TrtllmFP8 MoE 权重重排内存碎片化）：同属量化相关 bugfix，但解决不同问题。
  - PR #35733（支持 NVFP4 模型）：同属量化特性支持，涉及 compressed-tensors 格式，但本 PR 是 bugfix 而非新功能。
- 演进趋势：该 PR 反映了 vLLM 在支持新兴模型（如 Qwen3.5）和复杂量化配置时，需要不断优化底层权重加载逻辑，以处理融合层和 tuple 分区等高级特性。