PR #25974 完整报告

sgl-project/sglang

[Fix]: Restrict Kimi-K2.5 shared-experts fusion to Quark MXFP4 checkpoints

合并时间：2026-05-22 04:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25974

执行摘要

- 一句话：修复 Kimi-K2.5 共享专家融合对非 Quark 检查点的误启用
- 推荐动作：值得快速合入，因为它修复了一个导致标准 Kimi-K2.5 完全无法使用的严重回归。设计决策（将 quant_config.get_name() == "quark" 作为门控条件）合理且注释详尽。建议在后续工作中考虑增加对新量化格式的通用支持，或将此门控抽象为可扩展的“fusion-capable quant config 允许列表”。

功能与动机

修复夜间测试 TestKimiK25.test_kimi_k25 的持续失败（gsm8k 得分 0.000）。Bisection 定位到 PR #25390，该 PR 将 determine_num_fused_shared_experts 中的 n_routed_experts 允许列表从 {256} 扩至 {256, 384}，使得任何 Kimi-K2.5 加载都无条件启用共享专家融合。融合路径要求共享专家权重已预融合到 routed-experts 张量中，这仅对 amd/Kimi-K2.5-MXFP4（Quark）检查点为真；标准 moonshotai/Kimi-K2.5（compressed-tensors）检查点的共享专家存储在 quantization_config.ignore 中，融合加载器静默缺失所有 experts.wN_weight，导致输出为静默乱码。

实现拆解

该 PR 修改了一个文件（python/sglang/srt/models/deepseek_v2.py），通过在允许列表条件中增加对量化配置的检查，将 Kimi-K2.5 共享专家融合限制在 Quark MXFP4 检查点。

1. 文件与入口：变更位于 DeepseekV2ForCausalLM.determine_num_fused_shared_experts 方法（约第 2432 行）。该方法在模型初始化时被调用，根据配置和运行时条件决定是否启用共享专家融合（num_fused_shared_experts 设为 0 或大于 0 的值）。
2. 修改的核心逻辑：在原有的 allow-list 检查中，当 self.config.n_routed_experts == 384 时，额外要求 self.quant_config 不为 None 且 self.quant_config.get_name() == "quark"。具体地，在 or self.config.n_routed_experts not in (256, 384) or self.config.n_shared_experts != 1 之后，新增了一个 or 子句：(self.config.n_routed_experts == 384 and (self.quant_config is None or self.quant_config.get_name() != "quark"))。只有所有条件都通过（即不触发 disable_reason），融合才会被启用。
3. 变更效果：对于标准 moonshotai/Kimi-K2.5（quant_config 为 compressed-tensors 类型，get_name() 返回 "compressed_tensors"），条件 self.quant_config.get_name() != "

quark" 成立，因此 disable_reason 被设置为 "Config does not support fused shared expert(s).", 融合被禁用，恢复到安全的非融合路径。对于 amd/Kimi-K2.5-MXFP4（Quark），条件不成立，融合继续启用，保留了 #25390 带来的性能收益（+14.8% 输出吞吐量）。

4. 测试与验证：PR 提供了 8×H200 上的精度测试结果，显示修改后 moonshotai/Kimi-K2.5 的 GSM8K 得分从 0.000 恢复到 0.942（TP8）和 0.945（TP8+DP8），与回归前一致。加载日志中无 experts.wN_weight not found 警告，内存占用从异常的 84.39 GB 回落到正常的 72.19 GB。
5. 配套文件：无其他文件或配置变更。

关键文件：

- python/sglang/srt/models/deepseek_v2.py（模块 模型定义；类别 source；类型 data-contract）：唯一的变更文件，包含 determine_num_fused_shared_experts 方法的核心门控逻辑。修改控制了共享专家融合是否启用，直接决定了标准 Kimi-K2.5 加载是否正确。

关键符号：determine_num_fused_shared_experts

评论区精华

审查过程简洁，未出现实质性技术争议。

- gemini-code-assist[bot] 进行了代码审查，描述变更内容后表示“无反馈意见”。
- ch-wan（项目维护者）直接批准（“LGTM”）。

没有未解决的讨论或拒绝提议。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 风险极低，但有两点需关注：

1. 回归风险：对于 amd/Kimi-K2.5-MXFP4（Quark）检查点，门控逻辑确保融合继续启用（仅当 quant_config.get_name() == "quark" 时 disable_reason 不为 True），因此 #25390 带来的 AMD 性能收益得以保留。但若未来出现第三个量化格式（如 "mx_fp4"）且同样预融合了共享专家，此门控将错误地禁止融合。这部分可通过扩展条件或引入更通用的检查来缓解。
2. 兼容性风险：变更仅影响 n_routed_experts == 384 的路径，256（DeepSeek-V3/R1）路径完全不变。其他架构（如基于 DeepSeek-V2 的变体）不受影响。
3. 测试覆盖：PR 未添加新的单元测试来覆盖此门控逻辑。现有夜间回归测试（test_kimi_k25.py）验证了标准检查点的正确行为，但未测试 Quark 检查点或未来新格式。
   - 影响：
   - 用户影响：修复了标准 moonshotai/Kimi-K2.5 模型在 8 卡 H200 等 GPU 上精度为 0 的严重 bug。用户无需任何操作即可从回滚中获得恢复。AMD 上 amd/Kimi-K2.5-MXFP4 模型的性能提升不受影响。

- 系统影响：仅在加载标准 Kimi-K2.5 检查点时改变行为，加载时间和内存占用恢复正常（约 72 GB vs 84 GB）。对其他模型无影响。
- 团队影响：变更极简（+11/-1），审查和合并成本低。为未来处理类似检查点差异提供了模板和文档注释。
- 风险标记：缺少测试覆盖，回归修复

关联脉络

- PR #25390 [AMD] Enable shared-experts fusion with new KIMI-K2.5-MXFP4 model.: 本 PR 修复了 #25390 引入的回归。 #25390 将 n_routed_experts 允许列表从 {256} 扩至 {256, 384} 以支持 Kimi-K2.5-MXFP4 (Quark)，但未检查量化格式，导致标准 Kimi-K2.5 (compressed-tensors) 被错误地启用融合并产生乱码。本 PR 通过增加量化格式门控来限制该路径。