PR #25740 完整报告

sgl-project/sglang

[AMD] Bump amd/Kimi-K2.5-MXFP4 revision to align with shared-experts fusion

合并时间：2026-05-19 13:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25740

执行摘要

- 一句话：修复 Kimi-K2.5 MXFP4 测试因模型版本过旧失败
- 推荐动作：建议合并。该修复是同行 PR #25390 的配套变更，确保 AMD 路径的 Kimi-K2.5 测试不会因模型版本过旧而失败。

功能与动机

PR #25390 为 AMD 路径启用了 Kimi-K2.5 的 shared-experts fusion，但测试使用的旧 revision（b071bc6f）中 shared experts 是 bf16 未量化，导致 weight 加载时报 RuntimeError: The size of tensor a (3584) must match the size of tensor b (7168)。

实现拆解

1. 在 test/registered/amd/test_kimi_k25_mxfp4.py 中将 KIMI_K25_MXFP4_REVISION 从旧 commit SHA 更新为新的 HF main HEAD SHA，并添加多行注释说明升级原因及最低安全版本。
2. 已验证新 revision 的 shared experts 包含 180 个 MXFP4 weight_scale 条目，确保 dtype/shape 与 fusion 路径匹配。
3. 仅修改测试配置，无源码逻辑变更，属于最小风险修复。

关键文件：

- test/registered/amd/test_kimi_k25_mxfp4.py（模块 AMD 测试；类别 test；类型 test-coverage）：唯一变更文件，更新了模型 revision 常量以兼容 shared-experts fusion。

关键符号：未识别

关键源码片段

test/registered/amd/test_kimi_k25_mxfp4.py

唯一变更文件，更新了模型 revision 常量以兼容 shared-experts fusion。

```
# 测试文件：test/registered/amd/test_kimi_k25_mxfp4.py
# 原 revision b071bc6f 的 shared_experts 为 bf16（未量化）
# 新 revision 419004c8 将 shared_experts 量化至 MXFP4（uint8）
# 从而兼容 shared-experts fusion 路径（PR #25390）
# 最低安全版本为 94d8c1bd（HF 2026-04-01 修复）
```

```
KIMI_K25_MXFP4_REVISION = (
    "419004c8716cf22c929aa15d39b85e09a8a2091a"
)
```

评论区精华

该 PR 无 review 评论，但作者在 PR body 中详细对比了新旧 revision 的 weight_scale 条目数，并附上 CI 通过截图。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试中的模型版本号，不影响任何生产逻辑。但新 revision 可能引入模型行为的微小变化，已通过 CI 精度测试确认。
- 影响：影响范围仅限于 AMD CI 中 Kimi-K2.5-MXFP4 测试，修复后该测试能正常通过 shared-experts fusion 路径。
- 风险标记：依赖外部模型版本，低风险

关联脉络

- PR #25390 [AMD] Enable shared-experts fusion for Kimi-K2.5: 此 PR 是 #25390 的配套修复，确保其引入的 fusion 路径能正确加载 weight。
- PR #22188 [AMD] Pin Kimi-K2.5-MXFP4 revision: 该 PR 设置了旧的 revision pin，本次将其更新以兼容新功能。