PR #26774 完整报告

sgl-project/sglang

[NPU][Docs] Kimi-K2.5 best practice

合并时间：2026-06-02 13:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26774

执行摘要

- 一句话：为 Kimi-K2.5-w4a8 新增 Ascend NPU 最佳实践文档
- 推荐动作：值得精读，特别是需要在 NPU 上部署 Kimi K2.5 的用户。重点关注低延迟与高吞吐配置的差异，并注意表格与命令的卡数一致性。

功能与动机

Add best practice for Kimi-K2.5-w4a8 on NPU platform.

实现拆解

1. 新增 Kimi 系列最佳实践章节：在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx 末尾添加 "Kimi Series Models" 小节，包含低延迟和高吞吐两张配置表格，列出模型、硬件、卡数、部署模式、数据集、TPOT、量化方式等参数，并提供指向对应 Optimal Configuration 锚点的链接。
2. 修复占位符格式：在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_kimi_k2.5_examples.mdx 中将 Router 命令中的占位符从 'your prefill ip1' 等单引号格式改为 <your_prefill_ip1> 尖括号格式，增强可读性和准确性。
3. 冲突合并与格式修复：通过合并 main 分支解决冲突，并通过 commit "fix lint" 修复文档格式问题。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx（模块 NPU 最佳实践；类别 other；类型 documentation）：新增 Kimi 系列模型最佳实践章节，包含核心配置表格，是 PR 主要内容
- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_kimi_k2.5_examples.mdx（模块 NPU 示例；类别 other；类型 documentation）：修复 Router 命令中的占位符格式，提高准确性和可读性

关键符号：未识别

关键源码片段

docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_kimi_k2.5_examples.mdx

修复 Router 命令中的占位符格式，提高准确性和可读性

```
# 修改后的 Router 启动命令，使用尖括号占位符替换原单引号格式
python -m sglang_router.launch_router \
  --pd-disaggregation \
  --policy cache_aware \
  --prefill http://<your_prefill_ip1>:8000 8998 \
  --prefill http://<your_prefill_ip2>:8000 8999 \
  --prefill http://<your_prefill_ip3>:8000 9000 \
  --decode http://<your_decode_ip1>:8001 \
  --host 127.0.0.1 \
  --port 6688
# 注意：替换尖括号内的 IP 地址为实际 IP
```

评论区精华

Gemini Code Assist 机器人指出表格中卡数（8 卡）与 Optimal Configuration 命令中 --tp-size 16 不一致，并建议移除冗余的 transformers 安装命令。这些反馈未在后续 commit 中体现，PR 已合并但问题仍可能存在。

- 配置不一致及冗余建议 (correctness): PR 已合并，但表格卡片数问题未修复，可能需后续跟进。

风险与影响

- 风险：文档中表格的卡片数（8）与实际部署命令（--tp-size 16）不一致，可能导致用户误配置。此外，缺少多节点部署的详细说明可能增加调试成本。
- 影响：对用户：提供官方推荐配置，降低 Kimi-K2.5-w4a8 在 Ascend NPU 上的部署门槛。对系统：无运行时影响。对团队：需保持文档与实际参数一致性。
- 风险标记：文档配置不一致，用户部署失败风险

关联脉络

- 暂无明显关联 PR