PR #23077 完整报告

sgl-project/sglang

[NPU] [DOC] Update npu best practice docs to match latest code

合并时间：2026-04-18 14:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23077

执行摘要

本 PR 更新了 Ascend NPU 平台的最佳实践文档，同步了最新代码变更，包括调整 Deepseek-R1 的性能指标、添加 Qwen3 模型配置，并修正路由器启动命令的环境变量设置方式，以确保用户能基于优化后的配置进行部署。

功能与动机

动机是更新 Qwen3-8B、Qwen3-Next-80B-A3B 和 Deepseek-R1 的最佳实践文档，以匹配代码库的最新变化。PR body 中明确指出："Update best practice for qwen3-8b, qwen3-next-80b-a3b, deepseek-r1"，反映了文档需要与代码保持同步的需求。

实现拆解

1. 更新性能表格：在 docs/platforms/ascend/ascend_npu_best_practice.md 中，修改了多个表格数据。例如，将 Deepseek-R1 低延迟配置的 TPOT（每次输出时间）从 20ms 更新为 19ms，并添加了新的高吞吐量配置（如 24 卡部署）。 ```markdown 原表部分行，TPOT 从 20ms 改为 19ms，反映性能优化

| Deepseek-R1 | Atlas 800I A3 | 32 | PD Disaggregation | 3.9K+1K | 19ms | W8A8 INT8 | Optimal Configuration | ```

1. 添加新模型配置：为 Qwen3-8B 和 Qwen3-Next-80B-A3B 添加了详细的部署模式和配置参数，扩展了文档覆盖范围。
2. 标准化启动命令：路由器启动命令中，环境变量 SGLANG_DP_ROUND_ROBIN=1 最初设置为内联，后根据 review 反馈改为导出方式，提高稳健性。 shell export SGLANG_DP_ROUND_ROBIN=1 python -m sglang_router.launch_router \
3. 格式修正：提交历史显示，作者通过多个 commits（如"fix format"、"fix code review"）修复文档格式和响应审查意见，确保文档质量。
4. 测试和配置配套：无源码或测试改动，纯文档更新，因此无需额外配套。

评论区精华

review 中，gemini-code-assist[bot] 指出：

> "The environment variable SGLANG_DP_ROUND_ROBIN=1 is being set inline for a single command. If this variable is intended to be used by the router process, it is

> better to export it or set it in the environment configuration to ensure it is correctly picked up by the subprocesses spawned by the router."

作者采纳了建议，在后续提交中更新了命令格式，体现了对最佳实践的关注。

风险与影响

- 风险：文档错误可能导致用户配置不当，但因为是同步代码更新，风险较低；如果用户使用旧版本代码，新文档可能不兼容，但文档通常针对最新版本。
- 影响：对 NPU 用户，更新后的文档提供了准确的性能数据和配置指南，有助于提升部署效率；对团队，减少了因文档过时而产生的支持负担。

关联脉络

从近期历史 PR 看，PR 23009 也涉及 NPU 和文档更新，但侧重于代码移除。本 PR 是文档同步的一部分，反映了 NPU 平台持续优化的趋势。其他相关 PR 如性能优化可能间接影响文档内容，但无直接功能关联。