PR #25778 完整报告

sgl-project/sglang

[NPU] [DOC] remove Qwen3-235B-A22B 2K+2K 100ms mixed mode benchmark

合并时间：2026-05-19 20:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25778

PR 分析报告：[NPU] [DOC] remove Qwen3-235B-A22B 2K+2K 100ms mixed mode benchmark

执行摘要

此 PR 从 Ascend NPU 最佳实践文档中删除了 Qwen3-235B-A22B 模型在 Atlas 800I A3 硬件上的 2K+2K 100ms 混合模式基准配置，共删除 82 行，无新增代码。仅影响文档，无代码风险，但 reviewer 指出删除导致表格引用不一致，未被采纳。

功能与动机

根据 PR body，目的是移除该基准配置。可能原因包括该配置不再推荐、已被 50ms 基准替代，或存在错误。

实现拆解

1. 删除表格行：在 ascend_npu_best_practice.mdx 的基准汇总表中，移除了 Qwen3-235B-A22B / Atlas 800I A3 / 8 卡 / PD Mixed / 2K+2K / 100ms / W8A8 INT8 的对应行。
2. 删除详细配置节：移除了包括部署命令、环境变量设置和测试命令在内的完整配置节。

无源码变更。

评论区精华

reviewer gemini-code-assist[bot]指出：“删除 100ms 基准的表格行后，50ms 基准的表格行仍保留，导致不一致”，建议更新表格以引用 50ms 基准。该建议未被采纳，sglang-npu-bot 直接批准 PR。

风险与影响

- 风险：低。仅文档变更，但未修复 reviewer 指出的一致性问题，可能误导读者。
- 影响：小。仅影响参考 NPU 最佳实践文档的用户，移除的基准不再可用，但仍有 50ms 基准可参考。

关联脉络

此 PR 与 PR #25735 （[NPU] [DOCS] Improved the usability of Ascend NPU documents）同属 NPU 文档改进系列，后者重构了 NPU 文档结构，体现了团队对 NPU 文档持续优化的意

图。