PR #21807 完整报告

sgl-project/sglang

[NPU] update ascend docs

合并时间：2026-04-01 17:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21807

执行摘要

本次 PR 更新了 Ascend NPU 平台的文档，移除了已弃用参数并添加了 Qwen3-235B 模型在 256K 长序列上的运行示例，旨在提升文档准确性和覆盖范围。

功能与动机

PR 的动机是清理文档中已弃用的参数用法，并扩展文档以支持新模型和场景。具体来说：

- 移除已弃用的 SGLANG_DP_ROUND_ROBIN 等参数引用，避免用户配置错误。
- 为 Qwen3-235B-A22B 模型添加在 256K 长序列上运行的详细示例，帮助用户部署复杂推理任务。

实现拆解

实现涉及三个文档文件的修改：

1. *ascend_npu_best_practice.md*：移除了多处 *export* SGLANG_DP_ROUND_ROBIN=1 的引用。
2. *ascend_npu_deepseek_example.md*：类似地移除了该参数的引用。
3. *ascend_npu_qwen3_examples.md*：新增了以下内容：
   - 环境变量设置，如 ASCEND_USE_FIA=1 等。
   - 预填充节点、解码节点和路由器的详细启动命令，支持 Qwen3-235B-A22B 模型在 2 个 Atlas 800I A3 节点上运行 256K 长序列。

评论区精华

Review 过程中没有实质性讨论，仅有 sglang-npu-bot 的自动化批准，表明变更被直接接受。

风险与影响

- 风险：文档准确性是主要风险点。移除已弃用参数可能影响依赖这些参数的老用户，但鉴于参数已弃用，这是必要的清理。新增示例需要确保配置正确，避免误导用户。
- 影响：影响范围限于 Ascend NPU 平台的用户文档。移除已弃用参数提升文档清晰度；新增示例扩展了文档覆盖，支持更复杂的部署场景。

关联脉络

从历史 PR 看，本次文档更新与涉及 Ascend NPU 平台的代码变更（如 #17122）相关，但属于独立的文档维护工作，无直接代码关联。