PR #25268 完整报告

sgl-project/sglang

[NPU] [DOC] fix issues in ascend npu docs

合并时间：2026-05-14 17:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25268

执行摘要

本 PR 对 Ascend NPU 文档进行了全面修复，包括迁移 Docker 镜像地址、修正 MDX 渲染错误、添加推荐模型提示等。变更不涉及任何代码逻辑，风险极低，合并后能改善用户部署体验和文档可读性。

功能与动机

根据 PR 描述，动机有二：

1. 修复 Ascend NPU 文档中的 MDX 渲染问题；
2. 为某些场景补充解释说明。具体包括： - 将 quay.io/ascend/sglang 镜像仓库迁移至华为云 SWR，解决国内用户拉取镜像困难的问题（来自 commit 消息）。 - 将表格内 HTML 实体换行符
 替换为标准
，修复 MDX 渲染后多余空行或格式错乱。 - 为 Qwen3.5 部署示例添加推荐模型提示，帮助用户快速选择正确的模型权重（来自 commit 描述）。

实现拆解

1. 表格渲染修复（ascend_npu_support_features.mdx、ascend_npu_environment_variables.mdx、ascend_npu.mdx）：
   - 全局替换
 为
，确保 MDX 正确生成 HTML 换行。
   - 修正一处误删除单元格导致的表格列数不对齐问题（AI 审查发现）。
2. 镜像地址迁移（ascend_npu_quick_start.mdx、ascend_npu_qwen3_5_examples.mdx、ascend_npu.mdx）：
   - 将所有 quay.io/ascend/sglang:main-cann8.5.0-a3 等地址改为 swr.cn-southwest-2.myhuaweicloud.com/base_image/dockerhub/lmsysorg/sglang:main-cann8.5.0-a3。
3. 推荐模型提示（ascend_npu_qwen3_5_examples.mdx）：
   - 在 Qwen3.5 四个子章节中各插入一个 <Tip> 块，列出作者推荐的预训练量化模型名称。
4. 补充说明（ascend_npu.mdx）：
   - 在“获取 CANN 镜像”小节添加 <Note>，提醒用户 CANN 镜像与 SGLang 镜像托管在不同 registry。
   - 移除一行无用的 alias drun 语句，避免歧义。

所有修改均为纯文档内容，无测试或配置变更。

（本 PR 不涉及逻辑代码，聚焦文档 Markdown 修复，因此无代码片段展示。）

评论区精华

- gemini-code-assist[bot]在 ascend_npu_support_features.mdx:1059 发现：删除默认值单元格导致表格 4 列变为 3 列，要求恢复。作者在 commit fix AI code review 中修正，评论已标记为解决。

风险与影响

风险

- 镜像地址正确性：新地址若无效将导致拉取失败，但地址来源可靠，且为华为云官方 registry。
- 表格渲染兼容性：标准
 在绝大多数 Markdown 渲染器中支持，风险极低。
- 推荐模型名称：可能随模型版本更新而过时，但用户可自行调整。

影响

- 用户：受益于准确的镜像地址和推荐模型，降低部署门槛；文档表格渲染更清晰。
- 系统：无运行时影响。
- 团队：后续文档维护需保持同步。

关联脉络

本 PR 是独立文档修复，与历史 PR（如 #25130 #25050）无直接关联。但属于 NPU 文档持续改进的一部分，之前已有数次 NPU 文档调整（如 #23329）。整体趋势是完善 Ascend NPU 支持的用户文档，提高易用性。