# PR #25735 完整报告

sgl-project/sglang

[NPU] [DOCS] Improved the usability of Ascend NPU documents

合并时间：2026-05-19 16:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25735

## 执行摘要

本 PR 对 Ascend NPU 系列文档进行全面重构，包括统一镜像仓库地址、使用 Tabs 组件区分 A2/A3 硬件型号、新增磁盘空间警告、网络不可达 FAQ、服务测试指引等。所有变更为文档内容调整，无代码修改，显著提升了 NPU 文档的可读性和实用性。

## 功能与动机

根据 PR 描述，此变更旨在“Improved the usability of Ascend NPU documents”。原先的文档存在镜像地址分散（同时出现 swr.cn-southwest-2 和 quay.io）、未明确区分 A2/A3 配置（仅靠文字说明）、缺少常见问题排查指南等问题，导致用户在部署和测试 NPU 环境时容易混淆和出错。本次重构统一了镜像源，增加了结构化组件和警告提示，目标是为 NPU 用户提供清晰、一致、易于跟随的文档体验。

## 实现拆解

1. 统一镜像仓库和标签：将所有文档中引用的 Docker 镜像从 swr.cn-southwest-2.myhuaweicloud.com/base_image/dockerhub/lmsysorg/sglang 迁移至 quay.io/ascend/sglang，并区分稳定版（如 v0.5.10-npu.rc1-a3）和每日构建版（如 main-cann8.5.0-a3）两种标签，使用 Tip 提示用户选择。涉及 ascend_npu.mdx、ascend_npu_quick_start.mdx、示例文档等。
2. 使用 Tabs 组件区分硬件型号：将原先的纯文本说明“For Atlas 800I A2 users...”替换为 <Tabs> 组件，分别展示 Atlas 800I A3 和 A2 的完整命令。用户在文档中可以直接 Tab 切换，无需手动替换参数，大幅降低配置错误风险。
3. 增加磁盘空间警告和 FAQ：在所有镜像拉取和模型下载页面添加 <Warning> 组件，提示至少需要 30GB 空闲空间。在 FAQ 页新增网络错误（[Errno 101] Network is unreachable）的排查指南，提供了 HF 镜像代理、代理设置、手动下载三种方案。
4. 完善服务测试和端口配置：在 Qwen3.5 和 GLM5 示例文档中新增“Testing the Service”章节，包含验证命令。在主文档中增加端口使用建议，涵盖单机部署的内网映射和多节点部署的端口范围。
5. 链接和排版优化：将推荐模型名称升级为可点击的 ModelScope 链接，统一 “Not test yet” 为 “Not tested yet”，修复多节点部署的端口示例等细节问题。

（无代码变更，仅文档内容。）

## 评论区精华

此 PR 无实质性讨论，由 sglang-npu-bot 直接批准合并。所有变更为文档内容调整，无需代码审查。

## 风险与影响

- 风险：纯文档变更，无引入技术风险。镜像地址统一后，部分依赖旧地址的用户可能需更新脚本，但文档本身不存在执行风险。
- 影响：直接提升 Ascend NPU 用户的部署体验，新手可更快上手；FAQ 和测试指引有助于降低团队支持负担。预计对现有使用流程无破坏性影响。

## 关联脉络

此 PR 为独立的文档改进任务，未关联特定 Issue。与近期 NPU 相关的代码 PR（如 #22338 NPU MXFP4 量化、#24954 Mamba bugfix）没有直接依赖关系，但共同完善了 NPU 生态的文档和代码基础。后续建议继续补充更多模型的 NPU 部署示例。