

PR #28284 完整报告

sgl-project/sglang

Update documentation for Ascend NPU Guide

合并时间: 2026-06-15 20:02

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28284>

执行摘要

此 PR 更新 Ascend NPU 文档，统一镜像版本、添加离线下载方案、修正参数描述冲突，并优化格式。

功能与动机

根据 PR body，更新 NPU 文档以解决以下问题：

1. 网络下载问题，添加 HF 镜像和离线方案；
2. 统一快速开始中的镜像版本；
3. 修正 `--hcache-write-policy` 冲突描述；
4. 补全优化文档中的参数说明。

实现拆解

1. 统一 Docker 镜像版本：修改 `ascend_npu_quick_start.mdx`，将 Atlas 800I A3 和 910B 的稳定版本分别更新为 `v0.5.13.post1-cann9.0.0-a3` 和 `v0.5.13.post1-cann9.0.0-910b`，与每日构建版本保持统一格式。
2. 添加 Hugging Face 离线方案：修改 `ascend_npu_performance_testing.mdx`，在 Note 中新增 `--dataset-path` 离线方法，解决网络不可用问题。
3. 补全参数别名：修改 `ascend_npu_optimization.mdx`，为 `--dp-size` 和 `--ep` 补充别名。
4. 修正 HiCache 写策略描述：修改 `ascend_npu_support_features.mdx`，更新 `--hcache-write-policy` 的支持值。
5. 格式微调：根据 Review 处理换行和空格问题。

本次变更为纯文档修改，无源码变更。

评论区精华

- 版本一致性: `gemini-code-assist[bot]` 指出 `memfabric-hybrid` 安装命令更新到 1.0.8 但表格仍为 1.0.5，作者已同步更新。
- 列表渲染: `bot` 和 `amote-i` 均指出列表前缺少空行，作者已修复。
- 版本存在确认: `amote-i` 要求确认镜像版本存在，作者确认后合并。

风险与影响

低风险。主要风险是版本不一致和渲染问题，均已在审查中解决。影响范围为 NPU 用户，离线方案对特定网络环境有帮助。

关联脉络

此 PR 与 #28295 ([DOCS][NPU]Supplementary Notes) 同属 NPU 文档改进系列，逐步完善 NPU 指南。