PR #22307 完整报告

sgl-project/sglang

fix issues for npu docs

合并时间：2026-04-09 16:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22307

执行摘要

本次 PR 对 Ascend NPU 平台的文档进行了多处更新，主要修正了参数描述、版本信息和功能支持表格，旨在提升文档的准确性和可读性。变更仅限于文档内容，不涉及代码逻辑，因此风险极低，影响范围主要针对 NPU 平台用户和开发者。

功能与动机

根据 PR body 的描述，本次变更的动机是“为参数添加更清晰的选项描述”和“更新新的参数”。具体体现在文档中，例如：

- 为 --ep-dispatch-algorithm 参数补充了 static、dynamic、fake 等可选值。
- 更新了 CANN 工具包的版本号从 8.3.RC2 到 8.5.0，并同步了安装指南链接。
- 在功能支持表格中新增了 --enable-lora-overlap-loading 参数，并修正了 --moe-dense-tp-size 的默认值描述。

实现拆解

本次变更涉及 4 个文档文件，均为 Ascend NPU 平台相关： | 文件路径 | 主要变更 | 影响 | |----------|----------|------| | ascend_npu_support_features.md | 更新了多个参数的描述和选项，新增参数，调整支持表格 | 核心配置文件，影响用户参数理解 | | ascend_npu.md | 更新 CANN 版本号至 8.5.0 | 基础环境依赖指南 | | ascend_npu_quantization.md | 修正混合比特量化描述语句 | 量化功能文档 | | ascend_contribution_guide.md | 修正代码块格式 | 贡献指南格式优化 |

关键变更示例（来自 ascend_npu_support_features.md）：

```
- | `--pipeline-parallel-size`<br/>`--pp-size` | `1` | Type: int | A2, A3 |
+ | `--pipeline-parallel-size`<br/>`--pp-size` | `1` | Type: int; Currently `2` not supported | A2, A3 |
```

评论区精华

本次 PR 没有实质性的 review 讨论。唯一的 review 是由 sglang-npu-bot 自动批准的，没有提供具体意见。Issue 评论中仅有机器人关于配额限制的警告和触发 CI 的命令（/tag-and-rerun-ci）。因此，缺乏关于设计权衡或争议的讨论。

风险与影响

风险分析：

- 主要风险在于文档更新的准确性。例如，CANN 版本号或参数选项若与实际支持情况不符，可能误导用户。
- 由于是纯文档变更，不存在代码回归、性能、安全或兼容性风险。

影响分析：

- 对用户：提供了更准确的配置指南，减少配置错误，尤其明确了 --pp-size 当前不支持 2 等限制。
- 对系统：无直接影响。
- 对团队：提升文档质量，降低支持成本。

关联脉络

从近期历史 PR 看，NPU 平台的文档更新是一个持续的过程：

- PR #22429 同样更新了 NPU 文档，添加了 Qwen3 模型的低延迟配置指南。
- 本次 PR 中引用了历史 PR #17361（Advanced mix-bits for MoE）和 #14504（ModelSlim on Ascend support），表明文档在跟踪相关功能开发进展。 这反映了团队在不断完善 NPU 平台的文档体系，以支持该平台的特性演进和用户使用。