PR #22707 完整报告

sgl-project/sglang

[NPU] [DOC] Fix outdated descriptions in the NPU documentation

合并时间：2026-04-14 19:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22707

执行摘要

本次 PR 更新了 Ascend NPU 贡献指南文档，修复了三处过时描述：更正了准确性测试文件引用、澄清了 CI 冷却期适用对象、统一了硬件组件文件命名示例。这是一个低风险的文档维护变更，旨在提升贡献者体验，确保文档与代码实践同步。

功能与动机

PR 的动机明确为“修复 NPU 文档中的过时描述”。具体来说，文档中的以下内容已不符合当前项目状态：

- 引用的测试文件 test_moe_eval_accuracy_large.py 可能已不再使用或更名。
- CI 冷却期描述“Default cooldown period in minutes”未能明确针对低权限用户。
- 硬件组件文件命名示例 allocator_ascend.py 未使用更通用的“npu”后缀。

实现拆解

所有变更集中在单个文件 docs/platforms/ascend/ascend_contribution_guide.md，具体修改如下：

行号	原内容	新内容	目的
76	- [test_moe_eval_accuracy_large.py](...)	- [test_gpt_oss_1gpu.py](...)	更新准确性测试示例文件引用
116	description: "Default cooldown period in minutes; 0 disables rate limiting"	description: "Cooldown period in minutes for low-permission users; 0 disables rate limiting"	明确冷却期针对低权限用户
136	allocator_ascend.py	allocator_npu.py	使用通用 NPU 后缀统一文件命名示例

评论区精华

review 讨论非常简短，仅 gemini-code-assist[bot] 总结了变更要点：

> *“correcting a test file reference, clarifying the cooldown period description for low-permission users, and updating a file naming example to use a more generic NPU suffix.”*

该评论无进一步反馈，变更被快速批准。

风险与影响

- 风险：几乎为零。纯文档文本替换，不涉及代码逻辑、性能或安全。唯一潜在风险是引用的测试文件 test_gpt_oss_1gpu.py 可能在未来再次过时，但这属于常规文档维护范畴。
- 影响：直接影响 Ascend NPU 贡献者，确保文档指引准确，减少因过时信息导致的困惑。对系统运行时无任何影响。

关联脉络

- 与 PR #22793（修复 Ascend NPU 文档格式）类似，同属 NPU 文档维护序列，反映项目对硬件平台文档的持续更新。
- 文件命名示例从“ascend”改为“npu”，可能暗示项目在支持多 NPU 硬件（如 Ascend、AMD 等）时，倾向于使用通用后缀以保持一致性。