PR #26903 完整报告

sgl-project/sglang

[NPU] [DOC] clarify Ascend NPU exclusive supported values for speculative args

合并时间：2026-06-01 16:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26903

执行摘要

该 PR 仅修改了 Ascend NPU 支持特性文档中的两处表格描述，澄清了 --speculative-moe-a2a-backend 和 --speculative-draft-model-quantization 在 Ascend NPU 上的唯一支持值。变更极小，无代码或行为影响。

功能与动机

动机是明确文档以避免用户误用。PR body 原文："clarify Ascend NPU exclusive supported values for speculative args"。即告知用户，在 Ascend NPU 上 --speculative-moe-a2a-backend 仅支持 ascend_fuseep，--speculative-draft-model-quantization 仅支持 unquant。

实现拆解

1. 编辑 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_support_features.mdx 中的表格。
2. 在 --speculative-moe-a2a-backend 对应的值 ascend_fuseep 后添加 (the only supported value on Ascend NPU)。
3. 在 --speculative-draft-model-quantization 对应的值 unquant 后添加 (the only supported value for speculative decoding on Ascend NPU)。

无源码变更。

评论区精华

无实质讨论。仅 gemini-code-assist[bot] 自动确认无反馈。

风险与影响

- 风险：无。仅文档字符串修改，不影响软件行为。
- 影响：对 Ascend NPU 用户有积极指导作用，减少配置错误。

关联脉络

与之前 #26725（NPU MiniMax2.5 最佳实践文档）同属 Ascend NPU 文档改进系列，体现了对 NPU 平台文档持续完善的趋势。