

PR #29632 完整报告

sgl-project/sglang

[NPU] [DOC] Update deterministic inference feature support status to A2, A3

合并时间: 2026-06-29 19:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29632>

执行摘要

该 PR 将 Ascend NPU 功能特性文档中的 `--enable-deterministic-inference` 支持状态从 "Planned" 更新为 "A2, A3", 以反映该特性已在 A2 和 A3 平台上可用。这是一个单文件的文档变更, 无代码逻辑修改。

功能与动机

确保文档准确描述 `--enable-deterministic-inference` 特性在 Ascend NPU 上的实际支持情况, 避免用户混淆。

实现拆解

1. 在文件 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_support_features.mdx 中, 将特性支持表格内 `--enable-deterministic-inference` 对应的状态单元格从 "Planned" 修改为 "A2, A3"。

无关键源码。

评论区精华

无实质性讨论。

风险与影响

- 风险: 无。
- 影响: 提高文档准确性, 帮助用户正确配置确定性推理特性。影响面小。

关联脉络

该 PR 与历史 NPU 文档更新 PR (如 #29492) 属于同一文档维护系列, 持续完善 Ascend NPU 的支持说明。