

PR #30370 完整报告

sgl-project/sglang

[NPU] [DOC] Update features and mainstream models on ascend npu

合并时间: 2026-07-07 19:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30370>

执行摘要

此 PR 为纯文档维护, 更新了 Ascend NPU 平台的快速入门、支持模型列表和特性参数文档。添加了三个新模型 (DeepSeek-V4 变体、GLM-5.2), 替换了已废弃的参数说明, 并改进了文档间的导航链接。风险极低, 但对 NPU 平台用户具有实际指导价值。

功能与动机

根据 PR body, 动机是:

1. 更新 Ascend NPU 的支持特性与主流模型列表, 确保文档与最新代码同步。
2. 在快速入门页面添加 Note, 引导用户参阅完整的安装指南。

实现拆解

1. 快速入门页(ascend_npu_quick_start.mdx): 在页面顶部添加 <Note> 组件, 提示当前仅介绍官方容器镜像的部署方式, 完整安装指南 (源码安装、Docker 构建等) 需参阅另一文档。
2. 支持模型列表(ascend_npu_support_models.mdx): 在表格中插入三行新模型记录:
Eco-Tech/DeepSeek-V4-Pro-w4a8-mtp、Eco-Tech/DeepSeek-V4-Flash-w8a8-mtp、Eco-Tech/GLM-5.2-w8a8。同时修正 Eco-Tech/GLM-5.1-w4a8 的系列名从 "GLM-5.1" 纠正为 "GLM-5", 保持命名一致性。
3. 支持特性参数(ascend_npu_support_features.mdx): 新增 --enable-prefill-cp (bool 类型, A2/A3 支持) 和 --cp-strategy (zigzag/interleave 选项, A2/A3 支持) 两行参数说明。
删除已废弃的参数: --disable-piecewise-cuda-graph (建议使用 --cuda-graph-backend-prefill=disabled)、--enable-nsa-prefill-context-parallel 和 --enable-dsa-prefill-context-parallel (统一为 --enable-prefill-cp)、以及 --piecewise-cuda-graph-disable-attach-mlp (不再需要)。

(无代码变更, 仅文档 Markdown 文件修改)

评论区精华

- Gemini Code Assist 自动审查: 确认 PR 为文档更新, 无具体反馈。
- sglang-npu-bot 直接批准合并。

风险与影响

风险：极低。仅文档描述性变更，不涉及代码逻辑。潜在不一致风险（新参数描述与实际行为不符）已由维护者审阅控制。

影响：正面。使 NPU 平台文档更准确、更易导航，用户可快速了解支持的模型和正确的配置参数，避免使用废弃参数导致的错误。

关联脉络

- 与 #30328、#30358、#30324 等近期 NPU 文档 PR 共同构成 Ascend NPU 平台的持续文档改进。
- 模型列表的更新（特别是 DeepSeek-V4 变体）与仓库中多个 DeepSeek-V4 相关 bugfix（如 #30333）和性能优化（如 #30302）相对应，反映 NPU 对最新模型的适配进展。