

PR #30504 完整报告

sgl-project/sglang

[NPU] [DOC] Remove unsupported options of features on Ascend NPU

合并时间: 2026-07-08 17:35

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30504>

执行摘要

此 PR 更新了 Ascend NPU 的硬件平台文档，移除了 NPU 不支持的功能选项（如 `--enable-prefill-cp`、`--cp-strategy` 以及部分 CUDA Graph 后端）和不支持的模型（如 Moonshot Kimi Linear、PaddlePaddle ERNIE 4.5-VL），同时修正了部分 Qwen 模型名称。该变更由 NPU 维护者 sglang-npu-bot 批准，风险极低，主要是为了提升文档准确性和用户使用体验。

功能与动机

PR body 明确指出目的是移除 Ascend NPU 上不支持的功能选项。原文档中列出了 NPU 尚未支持或已放弃支持的功能和模型，容易误导用户尝试无效的配置。例如，`--enable-prefill-cp` 和 `--cp-strategy` 在 NPU 上并不可用；`--cuda-graph-backend-decode` 的 `breakable` 和 `tc_pieewise` 选项也不支持。本次清理旨在消除这些不一致。

实现拆解

- 更新 features 文档(ascend_npu_support_features.mdx):
 - 移除了 `--enable-prefill-cp` 和 `--cp-strategy` 两个选项。
 - 将 `--cuda-graph-backend-decode` 的可选值从 `full`, `breakable`, `tc_pieewise`, `disabled` 缩减为 `full`, `disabled`。
 - 将 `--cuda-graph-backend-prefill` 的可选值从 `breakable`, `tc_pieewise`, `disabled` 缩减为 `disabled`。
 - 将之前遗漏的 `--enable-prefill-cp` 条目以正确的顺序重新插入到文档表格中。
- 更新 models 文档(ascend_npu_support_models.mdx):
 - 移除了不再支持的两个模型行。
 - 修正了两个 Qwen 模型的名称，使其与官方命名一致。

整个过程仅涉及文档表格的行增删和文本修改，没有代码或配置变动。

无（纯文档变更）。

评论区精华

Gemini Code Assist 的自动评论仅总结变更内容，无额外反馈。sglang-npu-bot 直接批准。无技术讨论。

风险与影响

- 风险：极低。仅修改文档，不影响任何代码逻辑或系统行为。
- 影响：Ascend NPU 用户将获得更准确的参考文档，避免尝试不支持的功能；项目文档质量提升。

关联脉络

该 PR 是 Ascend NPU 平台文档的持续维护工作之一，与近期其他 NPU 相关 PR（如 #30496 AMD CI 修复）没有直接关联，属于独立的文档清理任务。