

PR #37799 完整报告

sgl-project/sglang

[NPU] [DOC] Refresh supported models and features on NPU

合并时间: 2026-09-03 19:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37799>

执行摘要

本次 PR 对 NPU（昇腾 NPU）平台的文档进行了维护性更新，主要移除了两个不再被支持的模型条目（FLM/Tele-FLM 和 CLIP），并将一个实验性功能（`--enable-attn-tp-input-scattered`）的状态描述统一为“Experimental”。此外，修复了文档示例中的几处微小拼写错误。这些变更是纯粹的文档更新，旨在提高信息的准确性和一致性，不涉及任何代码逻辑或功能变更。

功能与动机

动机源于保持 NPU 平台文档的时效性和准确性。PR body 中明确指出目标为“Refresh supported models and features on NPU”以及“Fix some docs issues on NPU”。通过移除过时或未实际支持的模型信息，并更正功能状态描述，可以避免用户在文档指引下进行无效配置，提升用户对 NPU 平台能力的认知准确性。

实现拆解

变更入口和过程非常直接：

1. 更新模型支持列表：核心变更是编辑 docs/docs/hardware-platforms/ascend-npus/reference/support_models.mdx 文件，从 Markdown 表格中删除了对应 FLM/Tele-FLM 和 AI-ModelScope/clip-vit-large-patch14-336 的整行数据。这直接修订了用户可参考的“数据契约”。
2. 修正功能状态说明：编辑 docs/docs/hardware-platforms/ascend-npus/reference/support_features.mdx，将 `--enable-attn-tp-input-scattered` 参数的“支持硬件”列从具体型号“A2, A3”改为“Experimental”。这修正了一个可能误导用户的配置信息。
3. 修复文档示例错误：在 profiling.mdx 和 operator_performance_optimizing.mdx 中，对日志输出和性能记录表格的示例内容进行了字符级别的修正（如增加空格、更正数字），属于纯文档打磨。

本次 PR 不涉及任何源码、测试或配置文件的修改，因此无需展示关键源码片段。

评论区精华

无讨论。PR 由自动化 bot `sglang-npu-bot` 直接批准，表明这是一个无需人工深度评审的常规文档维护任务。

风险与影响

- 风险：风险极低。所有变更均发生在文档文件（.mdx）中，不触及任何可执行代码、模型加载逻辑、内核或部署脚本。唯一潜在风险是文档更新后，其状态仍与未来代码实现脱节，但这属于持续维护范畴，非本次 PR 引入。
- 影响：影响限于文档读者。对于计划在 NPU 上运行 SGLang 的用户，本次更新提供了更干净、准确的模型和功能参考，有助于他们做出正确的技术选型和配置决策。

关联脉络

本次 PR 是 NPU 硬件平台持续文档维护的一部分。近期有其他 NPU 相关 PR，如 #37572（修复 NPU CI 任务名）和 #37193（XPU 模型启用），共同构成了对不同硬件后端生态支持的迭代工作。本次文档刷新是确保 NPU 平台文档与实现保持同步的常规环节。