

PR #29793 完整报告

sgl-project/sglang

[NPU]Modify --lora-backend & --moe-runner-backend description.

合并时间: 2026-07-01 11:15

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29793>

PR #29793 分析报告

执行摘要

该 PR 修正了 Ascend NPU 文档中 `--lora-backend` 和 `--moe-runner-backend` 参数的描述, 默认值和可选值均统一为 `ascend`, 并标注 `--moe-runner-backend` 仅适用于 GPU。变更极小, 仅涉及文档, 不影响源代码或运行时行为。

功能与动机

根据 PR body, 目的是“refine --lora-backend / --moe-runner-backend document”。原有文档中 `--lora-backend` 的默认值为 `csgmv`, 可选值包含 `triton`、`csgmv`、`ascend`、`torch_native`, 但实际 NPU 上仅支持 `ascend` 后端。同时 `--moe-runner-backend` 的支持硬件写为 "A2, A3", 但该参数实际仅用于 GPU。因此需要更新以反映真实情况, 避免用户困惑。

实现拆解

该 PR 仅修改了一个文件: `docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_support_features.mdx`。具体变更如下:

1. `--lora-backend` 行: 将默认值从 `csgmv` 改为 `ascend`, 可选值从多个后端精简为仅 `ascend`, 明确了 NPU 上只支持 `ascend` 后端。
2. `--moe-runner-backend` 行: 将“支持硬件”列从 "A2, A3" 改为 "Special for GPU", 表明该参数对 NPU 无意义, 仅用于 GPU。

无源码变更。

评论区精华

无 review 评论。

风险与影响

- 风险: 无。仅修改文档, 不影响任何代码或运行时行为。
- 影响: 影响范围极小, 仅对阅读 Ascend NPU 文档的用户有帮助, 防止误配置。

关联脉络

该 PR 是 NPU 文档持续完善的一部分, 类似于 PR #28586 对同一文档目录的格式与链接修正。同时, 与 AMD PR #27204 中间接涉及 `--moe-runner-backend` 硬件后端讨论相关。