

PR #21998 完整报告

sgl-project/sclang

[NPU] Optimized the wording in the npu docs

合并时间: 2026-04-03 11:51

原文链接: <http://prhub.com.cn/sgl-project/sclang/pull/21998>

执行摘要

本次 PR 优化了 NPU 相关文档的措辞，将“separation”统一替换为“disaggregation”并移除已不支持参数描述，以提升文档准确性和一致性，不影响代码逻辑。

功能与动机

动机源自统一术语和清理文档，具体为将“separation”替换为“disaggregation”以更准确描述部署模式（例如 PD 模式），并移除不再支持的 `--custom-sigquit-handler` 参数，确保文档与实现同步。PR body 中明确指出：“replace 'separation' with 'disaggregation'”和“remove unsupported argument”。

实现拆解

- `ascend_npu.md`: 修改场景标题从“PD Separation Scene”到“PD Disaggregation Scene”，调整一行代码。
- `ascend_npu_best_practice.md`: 在性能表格中批量替换“PD Separation”为“PD Disaggregation”，涉及多个配置项，确保术语统一。
- `ascend_npu_support_features.md`: 删除 `--custom-sigquit-handler` 参数行，清理过时信息。

评论区精华

无实质性讨论，仅由机器人 `sglang-npu-bot` 自动批准，表明变更直接且无争议，快速合并。

风险与影响

- 风险: 极低风险，主要在于术语替换可能遗漏其他文档实例（但 `patch_excerpt` 显示已全面替换），移除参数描述可能使用户误以为该参数仍存在（但正确清理）。
- 影响: 用户受益于更准确的文档；系统无任何影响；团队通过术语标准化提升协作效率。

关联脉络

与此前 NPU 相关 PR 如 #21408（支持 GLM-4.7-Flash）和 #21633（MOVA 扩散模型支持）呼应，是 NPU 功能文档维护的一部分，体现团队对文档质量的持续关注和 NPU 生态的演进。