PR #27032 完整报告

sgl-project/sglang

[NPU] add GLM model best practice docs

合并时间：2026-06-05 14:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27032

执行摘要

本 PR 在 Ascend NPU 最佳实践文档中新增了 GLM-5 和 GLM-5.1 模型的部署指南及基准测试数据，同时修复了 MiniMax-M2.5 锚点链接错误、硬编码用户路径等问题。纯文档变更，技术风险低，但对使用 NPU 部署 GLM 的用户具有直接参考价值。

功能与动机

PR 作者旨在为 GLM 模型补充在 Ascend NPU 上的最佳实践文档，使用户能够参考部署和调优。同时修复现有文档中的一些失效链接和硬编码路径，提升文档可用性。

实现拆解

1. 新增 GLM 模型基准测试章节：在 *ascend_npu_best_practice.mdx* 中添加了 GLM-5 和 GLM-5.1 的多个子章节，涵盖不同输入长度（3.5K+1.5K、64K+1K、128K+1K）和数据集（RANDOM、90% cache-hit）的场景。每个子章节包含模型标识、硬件规格、启动命令和性能数据表格。
2. 修复锚点链接：将 MiniMax-M2.5 和 GLM-5.1 的锚点 href 从错误的格式（如 #minimax-m2-5-...）修正为符合 slug 化规则的正确格式（如 #minimax-m25-...），避免跳转失效。
3. 替换硬编码路径：将部署命令中的 /home/luochen、/home/chenxu、/home/weights 等特定用户路径替换为通用的 /path/to/... 占位符。
4. 删除废弃参数：移除了示例中的 *--prefill-round-robin-balance* 参数。
5. 增加配置说明注释：为 SGLANG_SET_CPU_AFFINITY 等环境变量添加了适用场景注释，说明哪些选项仅适用于量化模型或非量化 MTP 层。

本 PR 为纯文档变更，不涉及代码，因此无需展示代码片段。

评论区精华

- 锚点链接 slug 化问题：*gemini-code-assist[bot]* 指出 MiniMax-M2.5 和 GLM-5.1 锚点中的点号被 slug 化去除，导致链接指向错误。作者最初认为无需修改，但最终提交中锚点被修正为正确格式。
- 硬编码路径问题：*gemini-code-assist[bot]* 指出命令中包含 /home/luochen 等用户特定路径，要求替换为占位符。作者全部采纳。
- 废弃参数移除：*amote-i* 指出 *--prefill-round-robin-balance* 已废弃，作者删除。

- 配置说明补充：cen121212 多次要求对 SGLANG_SET_CPU_AFFINITY、ZBAL 包、MTP 量化排除等添加注释，作者逐一增加了说明。

风险与影响

风险：

- 锚点链接仍可能因渲染引擎 slug 规则差异而失效。
- 性能数据基于特定版本，可能随时间过时。
- 命令中的环境变量可能因 Ascend 工具包版本变化而需调整。

影响：

- 用户：获得 GLM 模型在 NPU 上的详细部署指导，减少摸索成本。
- 系统：无直接影响。
- 团队：文档维护范围扩大至 GLM 系列，需与软件版本保持同步。

关联脉络

本 PR 与同仓库近期 #27308（文档同步）和 #27321（ cookbook 修复）同属文档维护工作，展示了团队在 NPU 文档上的持续投入。后续可能继续补充其他新模型的 NPU 最佳实践。