PR #26709 完整报告

sgl-project/sglang

[DOC] [NPU] add qwen3.5-397b best practice to doc_new

合并时间：2026-05-30 14:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26709

执行摘要

- 一句话：新增 Qwen3.5-397B 在昇腾 A3 的部署最佳实践
- 推荐动作：对于 Ascend NPU 用户值得阅读并参考其中配置；对于非 NPU 用户了解即可。文档组织方式和锚点链接设计可作为后续文档编写的参考。

功能与动机

PR 描述明确说明动机："Add Qwen3.5-397B-A17B low latency and high throughput best practicetonewdocument."即补充新模型在AscendNPU上的官方最佳实践，填补文档空白。

实现拆解

1. 在文档的模型配置总表中新增一行，列出 Qwen3.5-397B-A17B 的硬件、卡数、部署模式、长短上下文性能及量化方案，并链接到对应的详细配置章节。
2. 新增两个独立小节：" 低延迟配置 "（目标 22ms）和 " 高吞吐配置 "（目标 50ms），分别给出具体的启动命令、环境变量设置（如 SGLANG_NUM_PREFILL_TRITON_PAGE、SGLANG_BLOCK_SIZE 等）和 benchmark 命令。
3. 两个小节均包含模型名称、硬件、模式、数据集、输入输出长度、时延目标、基准测试命令等关键信息，并附带基准测试结果表格。
4. 文档全文为 .mdx 格式，使用 JSX table 组件，新增内容完全向后兼容。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx（模块 文档；类别 other；类型 docs）：唯一变更文件，新增 Qwen3.5-397B 低延迟和高吞吐两套部署配置及 benchmark 数据

关键符号：未识别

评论区精华

Review 中 gemini-code-assist[bot] 指出新增的锚点链接（如 #qwen35-397b-...）中包含了大写字母 B，而 Markdown 自动生成的标题 ID 均为小写，导致链接失效。建议将 #qwen35-397b-... 中 397B 改为 397b。该问题在最终合并的版本中可能已被修复（第二次 commit 为“fix lint”），但未见明确确认。

- 锚点链接大小写导致 Markdown 自动 ID 不匹配 (correctness): 问题被指出但未在评论区明确回复；PR 最终被合并，可能已通过后续 commit 修复。

风险与影响

- 风险：该 PR 仅涉及文档文件，风险极低。潜在风险包括锚点链接大小写不匹配（已通过 review 发现并可能已修正）和文档格式与已有风格不一致。对系统运行无任何影响。
- 影响：目标用户为使用 SGLang 在昇腾 NPU 上部署 Qwen3.5-397B 模型的开发者，提供可直接复用的配置和命令，降低调优成本。对现有功能无影响，文档体积增加约 200 行。
- 风险标记：文档链接正确性，格式一致性

关联脉络

- 暂无明显关联 PR