PR #25594 完整报告

sgl-project/sglang

[NPU] Add Qwen3.5-397B-A17B best practice doc

合并时间：2026-05-21 10:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25594

执行摘要

本 PR 为 Ascend NPU 平台新增 Qwen3.5-397B-A17B 模型的最佳实践文档，在性能表格中添加该模型条目，并提供了低延迟（22ms）和高吞吐量（50ms）两种场景的详细配置方案。文档现已合并，但存在模型名称不一致的 review 意见未被处理。

功能与动机

原 PR body 说明：'Add qwen3.5-397B best practice doc for Ascend NPU'。目的是为 NPU 用户提供已验证的大模型部署参数，降低调优成本。

实现拆解

1. 修改性能总表：在 'Low Latency' 和 'High Throughput' 两个表格中分别插入 Qwen3.5-397B-A17B 的行，记录硬件（Atlas 800I A3）、卡数（8）、部署模式（PD Mixed）、数据集规模、TPOT 延迟、量化方式（W4A8）及指向详细配置的锚点链接。
2. 新增配置章节：在文档末尾追加两个独立章节，标题如 Qwen3-397B 3_5K-1_5K 22ms on A3 8 Cards Mixed Mode（实际使用了简化名），包含模型全称、硬件清单、部署命令参数（如 --mm-attention-backend ascend_attn --dtype bfloat16 等）。
3. 格式修正：在后续 commit 中修复了 lint 问题，但未处理 review 提出的名称一致性与尾随空格问题。

本次变更为纯文档变更，未涉及程序源码。以下为文档中新增的性能表格行（Markdown 格式）：

| Qwen3.5-397B-A17B | Atlas 800I A3 | 8 | PD Mixed | 3.5K+1.5K | 22ms | W4A8 | [Optimal Configuration](#qwen3-397b-3_5k-1_5k-22ms-on-a3-8-cards-mixed-mode) |

但请注意，锚点链接中的模型名称为简化写法，与全称 Qwen3.5-397B-A17B 不一致。

评论区精华

- gemini-code-assist[bot]指出表格和章节标题使用了简化模型名 Qwen3-397B，而非正式名称 Qwen3.5-397B-A17B，建议统一使用全称且锚点全小写。
- gemini-code-assist[bot]还指出 shell 命令参数行存在尾随空格，违反代码风格规范。
- 作者未在讨论中回复，PR 即被 sglang-npu-bot批准合并。

风险与影响

- 文档准确性：模型名称不一致可能使用户混淆，锚点链接可能因大小写写错而失效。
- 可维护性：尾随空格虽不影响功能，但不符合项目 markdown 规范。
- 正面影响：提供了经过实测的配置参数，可大幅降低 NPU 用户部署 397B 大模型的尝试成本。

关联脉络

本 PR 与之前的 PR #23925（NPU 上为 Qwen3.5 引入融合 Triton 内核）属同一功能线：前者为模型提供运行时支持，此文档则帮助用户快速应用该支持，形成 ' 实现 + 部署指南 ' 的完整闭环。