PR #26725 完整报告

sgl-project/sglang

【NPU】add MiniMax2.5 best practice docs

合并时间：2026-06-01 10:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26725

执行摘要

- 一句话：为 NPU 新增 MiniMax2.5 最佳实践文档
- 推荐动作：文档清晰实用，建议合并。

功能与动机

提供 MiniMax2.5 模型在 NPU 上的最佳实践指导，便于用户复现性能和正确配置。

实现拆解

1. 在文档末尾新增 MiniMax Series Models 章节；
2. 提供低延迟配置表格，包含模型、硬件、卡数、部署模式、数据集长度、TPOT、量化方式和配置链接；
3. 后续段落（patch 后续部分）可能包含启动参数示例等详细配置。

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 documentation）：唯一变更文件，新增 MiniMax2.5 最佳实践内容

关键符号：未识别

关键源码片段

docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx

唯一变更文件，新增 MiniMax2.5 最佳实践内容

```
{/* 新增 MiniMax Series Models 章节开头 */}
## MiniMax Series Models

### Low Latency

{/* 性能配置表格 */}
<table style={{width: "100%", borderCollapse: "collapse", tableLayout: "fixed"}}>
  <colgroup>
    <col style={{width: "13%"}} />
```

<col style={{width: "13%"}} />
<col style={{width: "13%"}} />
<col style={{width: "13%"}} />
<col style={{width: "12%"}} />
<col style={{width: "12%"}} />
<col style={{width: "12%"}} />
<col style={{width: "12%"}} />
</colgroup>
<thead>
<tr style={{borderBottom: "2px solid #d55816"}}>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.02)"}}>Model</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.05)"}}>Hardware</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.02)"}}>Cards</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.05)"}}>Deploy Mode</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.02)"}}>Dataset</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.05)"}}>TPOT</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.02)"}}>Quantization</th>
<th style={{textAlign: "left", padding: "10px 12px", fontWeight: 700, whiteSpace: "nowrap", backgroundColor: "rgba(255,255,255,0.05)"}}>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiniMax-M2.5</td>
<td>Atlas 800I A3</td>
<td>8</td>
<td>PD Mixed</td>
<td>3.5K+1.5K</td>
<td>20ms</td>
<td>W8A8 INT8</td>
<td>Optimal Configuration</td>
</tr>
</tbody>
</table>

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：纯文档变更，无技术风险。
- 影响：影响使用 NPU 部署 MiniMax2.5 模型的用户，提供可参考的配置和性能预期。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR