PR #25875 完整报告

sgl-project/sglang

[NPU][DOCS]Add best practice and benchmark result parameter description

合并时间：2026-05-21 19:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25875

执行摘要

该 PR 为 Ascend NPU 文档新增了最佳实践指南与基准测试参数完整参考。主要变更包括在 ascend_npu_best_practice.mdx 中补充 Qwen3-27B 模型的 benchmark 数据行与详细测试结果章节，以及在 ascend_npu_performance_testing.mdx 中新增 SGLang Serving 基准测试输出参数的表格说明。Review 指出了多处输入输出长度标注与命令不一致的问题，但 PR 最终被批准合并。

功能与动机

提供更详尽的 NPU 部署参考和性能数据，帮助用户理解基准测试输出字段含义，便于调优与横向对比。文档面向 Ascend 用户，特别是使用 Atlas 800I A3 硬件的开发者。

实现拆解

1. 最佳实践文档扩展：在 ascend_npu_best_practice.mdx 的配置表格中新增 Qwen3-27B 的多行条目（覆盖不同卡数、精度、延迟场景），并在文档末尾增加 400+ 行 benchmark 结果章节，每个章节包含硬件配置、命令行、性能指标。
2. 基准测试参数参考：在 ascend_npu_performance_testing.mdx 末尾新增 "SGLang Serving Benchmark Result — Complete Reference"，用表格定义所有输出参数（如 Backend、Traffic request rate、Total input tokens 等）的说明与格式规范，帮助用户解读 bench_serving 输出。
3. 注释清理：移除 shell 命令示例中重复的 # bind cpu 注释。

最佳实践表格新增条目

```
// 在模型配置表格中新增 Qwen3-27B 条目，格式与其他模型一致
<tr>
  <td style={{padding: "9px 12px", fontWeight: 500}}>Qwen3-27B</td>
  <td style={{padding: "9px 12px"}}>Atlas 800I A3</td>
  <td style={{padding: "9px 12px"}}>2</td>
  <td style={{padding: "9px 12px"}}>W8A8 INT8</td>
  <td style={{padding: "9px 12px"}}>3.5K+1.5K</td>
  <td style={{padding: "9px 12px"}}>20ms (Mixed Mode)</td>
  <td style={{padding: "9px 12px"}}>
    <a href="#qwen3-27b-3_5k-1_5k-20ms-on-a3-2-cards-mixed-mode">Optimal Configuration</
    a>
  </td>
</tr>
```

```
// 注意：review 指出此处的 "3.5K+1.5K" 与实际命令顺序可能相反，应核对修正
```

基准测试输出参数参考

```
// 以下为表格一部分，列出核心统计字段及其格式
<table>
  <thead>
    <tr>
      <th>Parameter</th>
      <th>Description</th>
      <th>Format Specification</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td><code>Successful requests</code></td>
      <td>Total number of successfully completed requests (HTTP 200, no generation errors).</
      td>
      <td>Integer, no decimal places</td>
    </tr>
    <tr>
      <td><code>Benchmark duration (s)</code></td>
      <td>Total elapsed time from first request sent to last response fully received (seconds).</
      td>
      <td>2 decimal places</td>
    </tr>
    <tr>
      <td><code>Total input tokens</code></td>
      <td>Total number of input (prompt) tokens across all requests.</td>
      <td>Integer, no decimal places</td>
    </tr>
  </tbody>
</table>
```

评论区精华

- gemini-code-assist[bot]指出多处输入输出长度标注与命令不一致，例如标题 "16K+1K" 但命令实际为 1K 输入 + 16K 输出。该 bot 还发现 shell 命令中 # bind cpu 注释重复。所有评论均为 medium 优先级，作者未回复，但 PR 最终被批准，推测后续提交已修复。

风险与影响

纯文档变更，无代码风险。主要风险是数据不一致被修复前可能误导用户；但经 review 后应已解决。文档扩展降低了新用户的理解门槛，对 NPU 用户社区具有正面影响。

关联脉络

该 PR 与近期 NPU 相关 PR（如 #26466 软件升级、#26353 测试恢复）属于同一平台文档建设链路，共同提升 Ascend NPU 的文档完整性与测量标准。