PR #23845 完整报告

sgl-project/sglang

[Docs] Update Ascend NPU GGUF quantization documentation

合并时间：2026-04-27 22:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23845

执行摘要

本 PR 属于纯文档更新，主要针对 Ascend NPU 的 GGUF 量化支持进行文档同步，更新了量化支持表格、新增 GGUF 启动示例，并调整了相关支持标记。review 中发现示例缺少关键参数，但未修改即合并，存在误导用户的潜在风险。

功能与动机

随着 Ascend NPU 平台对 GGUF 量化模型的支持完善，需要更新文档以准确反映当前功能状态，并为用户提供可直接使用的启动命令。PR 旨在消除文档滞后，降低用户使用门槛。

实现拆解

1. 表格样式迁移：将 ascend_npu_quantization.mdx 中的 HTML span 内联样式全部替换为 React JSX 的 <strong style={{color: ...}}> 写法，统一文档工程风格。
2. 支持状态更新：在 quantization.mdx 中将 GGUF 对 NPU 的支持从“No”改为“Yes”，并注明实现方式为“CPU pre-dequantization at load time”。
3. 加载格式扩展：在 ascend_npu_support_features.mdx 的 --load-format 枚举中新增 gguf 值。
4. 示例添加：在 ascend_npu_quantization.mdx 末尾提供稠密模型和 MoE 模型的 GGUF 启动命令。

以下为 ascend_npu_quantization.mdx 中更新后的表格行示例，展示了 GGUF 支持标记和样式写法：

```
{/* 更新后的表格行示例：GGUF 支持状态，使用 React JSX 样式 */}
<tr>
  <td>GGUF</td>
  <td>Linear</td>
  <td><strong style={{color: 'green'}}>√</strong></td> {/* CUDA 支持 */}
  <td><strong style={{color: 'red'}}>x</strong></td>  {/* ROCm 不支持 */}
  <td><strong style={{color: 'green'}}>Yes</strong></td> {/* NPU：已支持，通过 CPU 预反量化 */
  }
</tr>
```

评论区精华

gemini-code-assist[bot] 指出两个启动示例均缺少 --load-format gguf 参数，并提供了修正版本。由于该建议未被采纳，当前文档中的命令可能无法直接执行。

风险与影响

- 风险：缺少 --load-format gguf 参数会导致加载失败，用户如果直接复制命令会遭遇错误，降低信任度。
- 影响：仅影响阅读文档的 NPU 用户，需关注后续是否有人提交修复 PR。建议团队尽快补充缺失参数。

关联脉络

该 PR 是 Ascend NPU 文档系列的一部分，与其他 NPU 相关文档（如 #23824 新模型支持指南）共同完善平台文档体系。后续建议与代码实现保持同步，确保文档示例经过测试。