PR #25384 完整报告

sgl-project/sglang

[NPU]Ascend NPU Performance Profiling Guide and Ascend NPU Operator Development Guide

合并时间：2026-05-21 17:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25384

执行摘要

- 一句话：添加 Ascend NPU 性能分析和算子开发两份文档
- 推荐动作：对 Ascend NPU 开发者或容器部署运维人员有较高参考价值，建议精读性能分析指南中的采集方法和算子开发指南的目录结构部分。其 review 讨论虽小但体现了文档术语准确性的重要性。

功能与动机

PR body 明确说明『add Ascend NPU Performance Profiling Guide and Ascend NPU Operator Development Guide』，旨在补齐 Ascend NPU 平台上关键文档缺失，帮助用户进行性能分析与自定义算子开发。

实现拆解

1. 编写 ascend_npu_profiling.mdx（+531 行）：阐述 SGLang 内置 PyTorch Profiler 在 Ascend NPU 上的使用，包括环境变量、四种启动 / 停止方式、trace 可视化以及注意事项；
2. 编写 ascend_npu_operator_development.mdx（+513 行）：介绍 SGL-Kernel-NPU 仓库结构，分步讲解 Ascend C 算子的 host/device 开发流程，并说明 Triton 算子的集成方式，附 HelloWorld 示例；
3. 更新 docs_new/docs.json（+2 行）：在 Ascend NPU 页面列表末尾添加两个新文档的引用路径，使其在导航栏中可见；
4. 更新 .codespellrc（+1/-1）：在 ignore-words-list 中加入 'CopyIn'，避免代码拼写检查误报此文档特有词汇。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_operator_development.mdx（模块 NPU 文档；类别 docs；类型 core-logic；符号 KernalHelloworld, TestHelloworld, test_helloworld_basic）：新增 513 行，系统讲解 Ascend C/Triton 算子在 SGLang 中的开发流程，是 NPU 自定义算子的核心参考文档。
- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_profiling.mdx（模块 NPU 文档；类别 docs；类型 dependency-wiring）：新增 531 行，详细说明如何利用 SGLang 内置 PyTorch Profiler 对 Ascend NPU 推理服务进行性能分析，覆盖环境变量、采集方法和可视化。

- docs_new/docs.json（模块 文档配置；类别 config；类型 configuration）：在 docs.json 的 Ascend NPU 页面组末尾添加两个新文档路径，使页面在导航栏中可访问。
- .codespellrc（模块 拼写配置；类别 other；类型 core-logic）：在 ignore-words-list 中添加 'CopyIn'，防止拼写检查将此文档特有用词报为错误。

关键符号：KernalHelloworld, TestHelloworld, test_helloworld_basic

评论区精华

来自 gemini-code-assist[bot] 的 review 评论：建议将 ascend_npu.mdx 中表格的『CANN Image Tag』标签改为『Hardware Identifier』，因为 910b、a3 实际为硬件标识而非完整 tag 名，与 Dockerfile 变量 DEVICE_TYPE 更一致。该建议未被合并前的对话解决，仍处于待处理状态。

- 表格标签命名建议：'CANN Image Tag' 改为 'Hardware Identifier' (style): 未在 PR 关闭前解决，评论状态仍为未解决。

风险与影响

- 风险：无代码级风险。拼写检查忽略列表新增词汇可能掩盖该词的真正拼写错误，但概率极低。文档内容与现有功能描述无冲突。
- 影响：直接影响 Ascend NPU 用户：提供性能分析工具使用指导和算子开发入门教程，可提升平台易用性。间接影响 SGLang 文档体系：补全硬件平台文档，降低 NPU 贡献门槛。影响范围局限于文档模块，无运行时影响。
- 风险标记：暂无

关联脉络

- PR #25257 [NPU] Support model DeepSeek-OCR and DeepSeek-OCR-2: 同为 Ascend NPU 平台功能 PR，本 PR 提供的算子开发指南可指导用户为 DeepSeek-OCR 开发自定义算子。
- PR #25839 [NPU] Support chunk prefill for Qwen3.5/Qwen3.6 models: 同为 Ascend NPU 平台功能 PR，本 PR 提供的性能分析指南可用于分析 chunk prefill 的性能瓶颈。