PR #23824 完整报告

sgl-project/sglang

[NPU] [DOC] Add support new models doc for NPU

合并时间：2026-04-27 17:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23824

执行摘要

- 一句话：为 NPU 新增新模型支持文档指南
- 推荐动作：建议 NPU 相关开发者和贡献者阅读此文档，以了解 SGLang 中支持新模型的标准流程和 NPU 适配要点。该文档整合了散落在各处的信息，是入门 NPU 模型支持的优质参考。

功能与动机

NPU 用户需要一份专门针对 Ascend 平台的模型支持指南，以了解如何在 SGLang 中适配新模型、利用 _is_npu 条件分支、调用 torch_npu 接口等。此前缺少此类文档，导致社区贡献者需要自行摸索。

实现拆解

1. 新增文档文件：在 docs_new/docs/hardware-platforms/ascend-npus/ 下创建 ascend_npu_support_new_models.mdx，共 551 行。文档从标题、描述、目录逐步展开。
2. 分章节讲解：包括“如何支持新语言模型”、“如何支持新多模态大模型”、“NPU 适配要点”、“从 vLLM 移植模型”、“注册外部模型实现”、“测试新模型”等章节。
3. 提供代码示例：包含 LlamaWrapper 封装示例、CustomQwen2VL 多模态模型示例、run_llm 入口函数等，并注明 NPU 特有改动。
4. 配套引用：链接到现有模型目录、配置文件和处理器目录，指导用户复用现有结构。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_support_new_models.mdx（模块 文档；类别 other；类型 data-contract；符号 import_new_model_classes, LlamaWrapper, init, forward）：唯一变更文件；提供了完整的 NPU 新模型支持文档，涵盖语言模型、多模态模型、NPU 适配、vLLM 迁移、外部注册和测试等内容。

关键符号：import_new_model_classes, LlamaWrapper, init, forward, main, run_llm, CustomQwen2VL, CustomQwen2VLProcessor

评论区精华

1. 硬编码行号问题：gemini-code-assist[bot] 指出文档链接 model_config.py 中使用了特定 commit 的行号 #L561，建议改用 main 分支无行号链接。作者回复“固定版本更容易找到”。

2. 排版错误：gemini-code-assist[bot] 发现测试章节尾部有多余的反斜杠 \，疑似 LaTeX 转换残留。
3. 函数签名缺失参数：CustomQwen2VL 的 forward 示例缺少 pp_proxy_tensors 参数，与 LlamaWrapper 示例不一致，可能导致运行时错误。作者回复“该参数是可选参数”。该问题在最终提交中已修复（fix AI review 提交）。

- 硬编码行号链接 (documentation): 作者回复 'A fixed version is easier to find.' 决定保留硬编码行号。
- 排版错误（多余反斜杠） (style): 已在后续提交中修复（fix AI review）。
- 代码示例缺少 pp_proxy_tensors 参数 (correctness): 作者回复 'the param is optional'，但为保持一致性，已在后续提交中补充该参数（fix AI review commit）。

风险与影响

- 风险：风险较低。主要风险是文档中的示例代码可能不准确或过时，如 forward 签名问题。但已通过 review 修正。此外，硬编码行号可能随代码变迁导致死链，但作者有意保留具体版本引用。总体风险可控。
- 影响：对 NPU 用户影响较大：提供了清晰的模型支持指南，降低了贡献门槛。对系统无影响，仅新增文档。对团队而言，完善了硬件平台文档体系，有利于社区贡献。
- 风险标记：文档示例可能过时，硬编码行号导致死链风险

关联脉络

- PR #23712 [Doc]Add msprobe doc in docs_new path: 同为 NPU 相关文档，扩展 docs_new 下硬件平台文档集。
- PR #20918 [NPU] Support MTP for Qwen3.5: 重要的 NPU 模型支持 PR，本 PR 文档可为类似贡献提供参考。