PR #24777 完整报告

sgl-project/sglang

[NPU] add Ascend NPU Accuracy Evaluation and Faq docs

合并时间：2026-05-14 11:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24777

执行摘要

- 一句话：新增 Ascend NPU 精度评估与 FAQ 文档
- 推荐动作：可直接合并。建议 NPU 用户精读精度评估文档以了解推荐流程，一线支持人员参考 FAQ 快速定位问题。后续可补充更多场景的 FAQ 条目。

功能与动机

根据 PR 描述，目的是为 NPU 用户提供标准化的精度评估流程指导和常见问题排障参考，降低 NPU 平台使用门槛。

实现拆解

1. 新增精度评估文档(ascend_npu_accuracy_evaluation.mdx)：详细说明环境搭建、使用 EvalScope 进行在线文本 / 多模态评估、使用 AISBench 进行离线评估，并提供命令示例。
2. 新增 FAQ 文档(ascend_npu_faq.mdx)：收集 PD 分离下的 context corruption 错误、graph 模式 acnn 错误、高并发长序列问题等，给出禁用 overlap 调度等解决方案，并附环境变量说明。
3. 更新导航配置(docs.json)：在 Ascend NPU 分组页面列表中插入两个新页面的路由。
4. 更新拼写忽略列表(.codespellrc)：添加单词 'tbe'，避免拼写检查误报。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_accuracy_evaluation.mdx（模块 文档；类别 other；类型 documentation）：新增精度评估文档，是 PR 核心内容之一，提供完整评估指南。
- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_faq.mdx（模块 文档；类别 other；类型 documentation；符号 alloc_extend）：新增 FAQ 文档，覆盖关键排障场景，直接帮助用户解决运行中问题。
- docs_new/docs.json（模块 配置；类别 config；类型 configuration）：更新导航配置，使两个新页面在文档站点中可访问。

关键符号：alloc_extend

评论区精华

无实质 review 讨论。PR 由 sglang-npu-bot 直接审批合并。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅涉及文档和配置文件，无代码变动。风险主要集中在文档内容的准确性，尤其是环境变量、命令参数和解决方案的正确性，若官方验证不充分可能误导用户。FAQ 中的临时方案（如禁用 overlap 调度）需注明已知问题状态。
- 影响：对 NPU 用户有正面影响：提供官方精度评估流程和常见问题排障路径，减少对支持团队的依赖。对系统其他模块无影响。文档导航更新确保用户可发现新内容。
- 风险标记：内容准确性依赖

关联脉络

- PR #25114 [NPU] [DOC] add performance testing and optimization docs for npu: 均为 NPU 文档系列，提供不同方面的 NPU 使用指南，共同完善 NPU 文档体系。