

PR #28279 完整报告

sgl-project/sglang

[NPU] fix ascend_docs

合并时间: 2026-06-15 21:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28279>

执行摘要

本 PR 对 Ascend NPU 四篇文档进行了维护性更新: 移除已废弃的 AISBench 评测工具、升级依赖版本、修正 EvalScope 命令参数、添加代理环境变量提示并调整算子开发示例。属于常规文档清理, 无源码变更, 风险低。

功能与动机

PR body 仅说明“fix ascend_docs”, 隐含目的是修正 Ascend NPU 文档中的过时信息和错误, 使其与当前推荐的 EvalScope 工具链及新版依赖保持一致, 减少用户困惑。

实现拆解

1. 精简 Accuracy Evaluation 文档 (ascend_npu_accuracy_evaluation.mdx) : 移除 AISBench 所有内容 (约 250 行), 只保留 EvalScope; 将 `--eval-type server` 改为 `--eval-type openai_api`; 数据集名 `math` → `math_500`、`mmbench` → `mm_bench`; 新增代理 `unset` 的 Warning 块。
2. 升级依赖版本 (ascend_npu.mdx) : MemFabric-Hybrid 1.0.5 → 1.0.8, PyTorch 2.8.0 → 2.10.0, TorchVision 0.23.0 → 0.25.0, TorchNPU 2.8.0.post2 → 2.10.0, triton-ascend 固定为 3.2.1。
3. 修正 FAQ 代码缩进 (ascend_npu_faq.mdx) : 调整 `out_indices` 示例的缩进。
4. 调整算子开发测试 (ascend_npu_operator_development.mdx) : 张量大小 1024 → 2048, 数据类型 `bfloat16` → `float16`。

因本 PR 无代码文件变更, 仅涉及文档 Markdown, 故不提供源码片段。

评论区精华

- gemini-code-assist[bot]指出描述仍写“two tools: EvalScope”, 应改为仅提及 EvalScope。该评论被提交后, PR 由 sglang-npu-bot 直接 approve, 未见进一步讨论。建议确认该处是否已修正。

风险与影响

- 风险: 极低, 仅文档变更。但依赖版本号若与用户环境不匹配可能造成安装困难; 命令参数 `openai_api` 和数据集 `math_500` 需与 EvalScope 实际版本兼容。
- 影响: 用户参考文档时获得更准确、最新的指导。移除 AISBench 可避免用户尝试已废弃工具。无系统运行影响。

关联脉络

本 PR 与 #28283、#28284、#28295 等同期的 NPU 文档 PR 共同构成了对 Ascend NPU 文档的集中清理与升级，反映团队正在主动维护 NPU 平台文档的实时性。