PR #24668 完整报告

sgl-project/sglang

[NPU]Documentation update for communications quantization feature

合并时间：2026-05-11 04:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24668

执行摘要

为 NPU 后端通信量化特性 --enable-quant-communications 补充文档描述，在 server_arguments.mdx 参数表格中新增一行说明，帮助用户了解该参数的作用与适用场景。变更简单，无代码或测试影响，已由 reviewer 批准合入。

功能与动机

PR #20520 为 NPU 后端引入了 TP 通信 INT8 量化特性（--enable-quant-communications），但缺少用户文档。本 PR 旨在补充该参数的描述，让使用者能清楚了解其功能（启用 TP 通信 INT8 量化）和限制（仅支持 NPU + Qwen3 系列），以正确配置和使用。

实现拆解

1. 定位变更文件：docs_new/docs/advanced_features/server_arguments.mdx，该文件列出了 SGLang 服务端的所有可选参数。
2. 新增文档行：在参数表格末尾（位于 --disable-cuda-graph-padding 之后）插入一行，包含参数名 --enable-quant-communications、类型 bool flag、默认值 False、说明文字。
3. 说明文字：明确标注该特性仅适用于 NPU 平台且仅限 Qwen3 系列模型，避免用户在其他环境下误用。

整个变更仅涉及文档表格添加一行，无代码、测试或配置改动。

本次 PR 只修改文档，无关键源码变更。

评论区精华

无实质讨论。PR 由 sglang-npu-bot 合入，审核人 ping1jing2 直接批准，未产生评论。

风险与影响

- 风险：无。仅纯文档修改，不影响任何代码逻辑或系统行为。
- 影响：对用户而言，补充了 NPU 通信量化特性的使用说明，有助于功能推广；对开发团队则是一次标准的文档补录。

关联脉络

- 关联 PR #20520：本 PR 的文档变更是对 #20520（引入 NPU 通信量化特性）的配套收尾工作。

- 无其他跨 PR 关联。