# PR #23564 完整报告

sgl-project/sglang

[NPU] [DOC] Update supported models and features of npu

合并时间：2026-04-25 15:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23564

## 执行摘要

PR #23564 更新了 Ascend NPU 平台的官方文档，补充了 SSL/TLS、Diffusion LLM 等服务器参数说明，并扩展了支持模型列表，新增 Qwen3.5-397B、Kimi-Linear-48B 等模型。变更不涉及任何代码或配置，风险极低，但对 NPU 用户有直接参考价值。

## 功能与动机

PR 标题与描述明确指出本修改的目的是“更新 NPU 支持的模型和特性”。随着 NPU 后端持续演进，原有文档已落后于实际支持能力，因此需要同步更新以确保用户能获取准确信息。

## 实现拆解

1. 更新特性文档（ascend_npu_support_features.mdx） - 新增 SSL/TLS 章节，列出 --ssl-keyfile、--ssl-certfile 等 5 个参数及其默认值、类型和适用平台。 - 新增 Diffusion LLM 章节，涵盖扩散模型相关参数。 - 其他已有参数表格可能也做了增量补充（diff 显示共 224 行新增）。
2. 更新模型支持文档（ascend_npu_support_models.mdx） - 在模型支持表格中插入多行新模型，包括：
   - Qwen/Qwen3.5-397B-A17B（Qwen 系列）
   - moonshotai/Kimi-Linear-48B-A3B-Instruct（Kimi Linear）
   - FLM/Tele-FLM（Tele FLM） 等，并标注“编译”和“推理”均受支持（✔）。
3. 格式与细节调整 - Review 中指出的表格 <colgroup> 列数不匹配问题以及参数默认值笔误，在合并前未修复，但未影响合并决策。

## 评论区精华

自动化助手 gemini-code-assist[bot] 提出了 3 条 medium 优先级的问题：

> “SSL/TLS 表格 colgroup 定义了 5 列，但表头只有 4 列，建议调整为 4 列各 25%。” “Diffusion LLM 表格也存在同样问题。” “--enforce-shared-experts-fusion 的默认值在文档中写为 True，但代码中是 False，可能是笔误。”

这些评论未被作者回复或修正，但最终由 sglang-npu-bot 直接批准合并。

## 风险与影响

风险：纯文档变更，无代码风险。但表格格式问题和默认值错误未修复，可能对读者造成轻微误导，影响有限。

影响：NPU 用户可获得最新的支持信息，有利于快速决策。此外，文档维护压力略有增加，需持续跟踪 NPU 进展。

## 关联脉络

本 PR 与近期 NPU 相关的代码变更（如 #23671、#23642）无直接关联，属于独立的文档同步工作。配合仓库中其他 NPU 演进 PR，反映了 SGLang 对 Ascend NPU 生态的持续投入。