PR #24676 完整报告

sgl-project/sglang

[NPU] [DOC] refresh npu supported model list

合并时间：2026-05-08 17:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24676

执行摘要

本次 PR 刷新了 Ascend NPU 支持的模型列表文档，新增了多款 Qwen3.5、Qwen3.6、GLM-5、Kimi 和 MiniMax 等模型条目，并删除了部分旧模型。变更仅涉及文档，不包含代码修改，属于常规维护。

功能与动机

根据 PR 描述，目的是 "refresh npu supported model list"，即更新 NPU 平台支持的模型列表，确保用户参考的文档与实际兼容性保持一致。

实现拆解

1. 识别新增模型：从 Eco-Tech 等来源确认了新支持的模型，包括量化的 Qwen3.5、Qwen3.6、GLM-5、Kimi-K2.6、MiniMax 等。
2. 修改文档表格：在 ascend_npu_support_models.mdx 的 <table> 中为每个新模型添加独立的 <tr> 行，填写 Model ID、Family、Offline/Online 支持状态。
3. 移除旧模型：删除不再支持的条目，如 Qwen/Qwen3.5-397B-A17B。
4. 处理 Review 反馈：针对机器人评论中的家族命名一致性和版本匹配问题，作者未修改文档，第二个问题回复 "Non-issue" 并合并。

以下为新增模型行的示例，注意 Model Family 列的值与 reviewer 建议的 'Qwen' 不一致：

```
<tr>
  <td>Eco-Tech/Qwen3.6-35B-A3B-w8a8</td>
  <!-- Model Family 使用了 Qwen3.6，与 reviewer 建议的 Qwen 不一致，但保留原样 -->
  <td>Qwen3.6</td>
  <td>✔</td>
  <td>✔</td>
</tr>
```

被删除的旧模型行格式如下： <!-- 删除的行：<td>Qwen/Qwen3.5-397B-A17B</td> -->

评论区精华

- gemini-code-assist[bot] 提出 Qwen3.6 的 Model Family 应写 'Qwen' 而非 'Qwen3.6'，但未获 author 回应，维持原状。

- 同样 bot 指出 Kimi-K2.6 在代码中仅有 K2.5 配置，可能不匹配。author amote-i 回复 "Non-issue"，认为版本号准确，最终合并。

风险与影响

- 风险：文档与代码版本可能脱节，如 Kimi-K2.6 和 Qwen3.6 在代码中尚无对应配置，可能误导用户。建议在文档更新时同步验证代码支持。
- 影响：用户获得最新支持列表，减少尝试不兼容模型的时间；对系统无影响；团队需持续维护文档与代码对齐。

关联脉络

- 本 PR 与 #24658（修复同一文档的目录）和 #23708（启用 GLM-5 文档的 DeepEP）同属 NPU 文档改进系列，共同完善 Ascend 平台的文档质量。