

PR #39555 完整报告

sgl-project/sglang

[NPU] [DOC] delete unsupported models in npu docs

合并时间: 2026-09-15 14:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39555>

[NPU] [DOC] 删除 NPU 文档中不受支持的模型

执行摘要

本 PR 是纯文档清理: 在 `docs/docs/hardware-platforms/ascend-npus/reference/support_models.mdx` 中整行删除 4 个模型条目 (共 -24 行、+0 行), 把 Ascend NPU 支持模型矩阵里「标着 ④ 但实际不受支持」的行摘掉。改动不触及任何源码、kernel、测试或配置, 运行时零影响; 唯一后果是 NPU 用户的选型参考更贴近真实能力, 代价是删除理由缺少可追溯记录。

功能与动机

PR 标题与 body 只有一句表述:

```
delete unsupported models in npu docs
```

也就是说, Ascend NPU 支持矩阵中列出的部分模型在 NPU 后端上并不受支持, 勾选状态与真实能力不符, 继续保留会误导部署选型。关联 Issue 列表为空, PR body 的 Modifications / Accuracy Tests / Speed Tests 三节均为 N/A, Checklist 也全部未勾选——作者明确把这次改动定位为事实性纠偏, 而非功能或性能变更。

需要标注的不确定性: 材料中没有 Issue 或评论解释「为什么这几个模型不受支持」, 因此是特定 CANN / torch-npu 版本限制、缺少算子实现, 还是仅仅没有验证过, 无法确认。

实现拆解

变更入口只有一处: `support_models.mdx` 这一个 .mdx 文件, 两处删除动作, 合计 24 行。

- 删除点一 (文件中段): 整行移除 LFM2 系列两行——LiquidAI/LFM2.5-1.2B-Instruct (系列列写作 LFM2 (230M, 350M, 1.2B)) 与 LiquidAI/LFM2-8B-A1B (LFM2-MoE (8B-A1B, 24B-A2B)), 两行原本都是 ④ / ④。删除后, 原先紧随其后的 `ibm-granite/granite-4.0-micro` 直接承接上下文。
- 删除点二 (文件后段): 同一处动作里移除 `AI-ModelScope/whisper-large-v3` (Whisper, 语音分组) 与 `LiquidAI/LFM2.5-VL-1.6B` (LFM2-VL, 多模态分组); 夹在两者之间的 `Qwen/Qwen3-ASR-1.7B` 被保留。也就是说语音分组删完后只剩 `Qwen3-ASR`, 多模态分组条目继续保留在表尾。

被删除行	系列	分组（依据相邻行推断）	删除位置
LiquidAI/LFM2.5-1.2B-Instruct	LFM2 (230M, 350M, 1.2B)	通用 LLM 分组	文件中段
LiquidAI/LFM2-8B-A1B	LFM2-MoE (8B-A1B, 24B-A2B)	通用 LLM 分组	文件中段
AI-ModelScope/whisper-large-v3	Whisper	语音 / ASR 分组	文件后段
LiquidAI/LFM2.5-VL-1.6B	LFM2-VL	多模态分组	文件后段

- 1. 表格契约保持不变：每行仍是四个 td（模型仓库 / 系列 / 两列硬件可用性），内联 style（padding: "9px 12px"、rgba(255,255,255,0.02|0.05) 背景）逐字未改。因此 Ascend A2 / A3 两列的勾选语义、表头与排序规则都没有变化，读者看到的只是行数减少——这种「只删行、不改 schema」的做法有效规避了文档渲染层的回归风险。
- 2. 判定粒度是单个模型：删除点二里，Whisper 与 LFM2.5-VL 被移除而中间的 Qwen3-ASR 保留，说明作者是按「单模型是否被 NPU 支持」逐行判断，而不是整段砍掉某个系列或某个模态。
- 3. 测试 / 配置 / 部署配套：无。没有新增测试、配置项或部署脚本改动。PR body 内嵌 CI 状态显示 Latest PR Test (Base) 通过、Latest PR Test (Extra) 在 Run #34937749583 失败，AMD ROCm 无运行记录；对只改 .mdx 的 PR，Base 通过已足以支撑合并，Extra 失败更可能是既有的不稳定任务，但这一点在材料中无法证实。

以下是根据删除点两侧上下文整理的表格片段，用于说明该支持矩阵的契约形态，以及本次删除发生在哪些行之间：

```
{/* docs/docs/hardware-platforms/ascend-npus/reference/support_models.mdx
Ascend NPU 支持模型矩阵：<tbody> 中每行代表一个模型条目，
四个 td 依次为 模型仓库 / 系列 / 两列硬件可用性勾选。
本次 PR 只做删行，表结构与内联 style 一律不动。 */}

<tbody>
  {/* 删除点 1：LFM2 系列两行被整行摘掉
    - LiquidAI/LFM2.5-1.2B-Instruct（系列 LFM2 230M / 350M / 1.2B）
    - LiquidAI/LFM2-8B-A1B（系列 LFM2-MoE 8B-A1B / 24B-A2B）
    两行原本都是 ❷ / ❸，按 PR 描述它们在 NPU 上其实不受支持，
    保留会直接误导部署选型；删除后 granite-4.0-micro 成为该处相邻行。 */}
  <tr>
    <td style={{padding: "9px 12px", fontWeight: 500, backgroundColor: "rgba(255,255,255,0.02)"}}>ibm-granite/granite-4.0-micro</td>
    <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.05)"}}>IBM Granite 4.0 (Hybrid, Dense)</td>
    <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.02)"}}>❷</td>
    <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.05)"}}>❸</td>
  </tr>
```

```

/* 语音分组：AI-ModelScope/whisper-large-v3 被移除后，
   该分组只剩下这一行 Qwen3-ASR；夹在中间的条目被保留，
   说明判定粒度是「单个模型是否被 NPU 支持」，
   而不是整段砍掉某个系列或某个模态。 */
<tr>
  <td style={{padding: "9px 12px", fontWeight: 500, backgroundColor: "rgba(255,255,255,0.02)"}}>Qwen/Qwen3-ASR-1.7B</td>
  <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.05)"}}>Qwen3-ASR (0.6B, 1.7B)</td>
  <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.02)"}}>🔒</td>
  <td style={{padding: "9px 12px", backgroundColor: "rgba(255,255,255,0.05)"}}>🔒</td>
</tr>

/* 多模态分组：紧随 Qwen3-ASR 之后、</tbody> 之前的
   LiquidAI/LFM2.5-VL-1.6B (LFM2-VL) 也在本次被整行删除，
   与 Whisper 同属后段这一处删除；表格列数与 style 完全未变。 */
</tbody>

```

评论区精华

本 PR 没有任何 review 评论（`review_comments_count = 0`），也没有 Issue 讨论，唯一的审核记录是机器人：

sglang-npu-bot: **APPROVED** (review body 为空)

因此不存在被讨论、被质疑或需要权衡的技术分歧。反过来看，也没有任何人追问「这几个模型为什么不受支持」「删除依据是什么」——判定依据只存在于删除动作本身。

风险与影响

风险项	说明	严重度
判定依据不可追溯	4 个模型行没有任何说明性文字（如「暂不支持」「未验证」），读者只能靠 git 历史反查消失原因；若某模型只在特定 CANN / torch-npu 版本或特定 A2/A3 形态下可用，这次删除会把信息抹平	中
文档与实现反向漂移	<code>support_models.mdx</code> 没有生成或校验机制。本次是单向收缩，未来 NPU 侧补齐 LFM2 / Whisper / LFM2-VL 支持后文档不会自动回补，容易出现「支持但文档没写」	中
整表未同步复核	两处删除共 24 行，但同一张表内其他模型行的勾选值没有一并复核；如果判定标准是「NPU 后端是否真跑通」，其余行可能仍有同类不实勾选	低
运行时回归	不涉及源码、kernel、配置或测试，运行时行为、性能、兼容性零影响	无

影响范围：仅限 Ascend NPU 部署者的选型路径与文档可信度。原先按矩阵去尝试 LFM2 系列、LFM2.5-VL 或 Whisper 的用户会少走弯路；但对系统调度、KV cache、attention、模型加载等模块均无影响。

关联脉络

- PR#39389 [NPU][DOC] Rename NPU hardware to Ascend A2/A3 Series product: 改动同一 docs/docs/hardware-platforms/ascend-npus/ 文档目录，把硬件称谓统一为 Ascend A2/A3 系列。本 PR 支持矩阵中的两列勾选正对应这两代硬件，二者同属一条「NPU 文档校准」脉络：先统一硬件称谓，再收缩不受支持的模型条目。
- PR#39423 [NPU][Bugfix] Disable pinned memory to fix DeepSeek-V2 DP-attention hang: 同为近期 NPU 平台方向变更，涉及 python/sglang/srt/platforms/npu.py。这类后端行为调整正是支持矩阵容易与实现脱节的根源，也解释了为什么该文档需要人工持续校正。
- 本次未引用任何 Issue，也没有 follow-up。若要把文档可信度做成机制而不是人工维护，可考虑为支持矩阵引入状态标记（如「已验证 / 社区贡献 / 不支持」），让后续删改自带理由，而不只是行数变化。