PR #35836 完整报告

sgl-project/sglang

[NPU] [DOC] Refresh supported features and models on Ascend NPU

合并时间：2026-08-23 13:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35836

PR 35836 分析报告：Ascend NPU 支持矩阵刷新

执行摘要

纯文档 PR：刷新 Ascend NPU 平台的参数支持状态与模型支持清单，将一批 Planned/Experimental/Special for GPU 参数升级为 A2/A3 可用，新增十余个模型行并移除 grok-2。变更不涉及代码，但两张表格是 NPU 用户选型与参数判断的直接依据，信息量较大且影响面明确。

功能与动机

PR body 仅一句话：Refresh supported features and models on Ascend NPU。结合提交拆解可见，NPU 后端近期的能力扩展（调度策略、MoE 数据并行、流水线并行、SWA 混合缓存、LoRA 严格加载、多模态模型等）大多已完成适配，但文档仍停留在旧的能力边界。刷新后，用户不会再因为文档标记为 Planned 而误判这些能力在 NPU 上不可用。

实现拆解

1. 参数状态升级（support_features.mdx）：将 --swa-full-tokens-ratio、--disable-hybrid-swa-memory、--enable-dynamic-chunking、--moe-data-parallel-size、--pp-max-micro-batch-size、--pp-async-batch-depth、--model-checksum、--lora-strict-loading 等参数从 Planned/Experimental/Special for GPU 改为 "A2, A3" 支持状态。其中 --lora-strict-loading 从 "Special for GPU" 变为通用支持，暗示 NPU 上 LoRA 能力已补齐。
2. 新增参数行（support_features.mdx）：补充 --retraction-policy（取值 length/priority）、--enable-session-radix-cache、--default-chat-template-kwargs 三行，并修正 --pp-async-batch-depth 默认值从 None 更正为 0，消除表格内不一致。
3. 模型清单刷新（support_models.mdx）：新增 Qwen3.8-2.4T-A95B、sgl-npu/Kimi-K3-W4A8（NPU 社区量化变体）、MiMo-V2-Flash-W8A8、LiquidAI LFM2 系列（1.2B 与 8B-A1B MoE）、Falcon-H1-34B、IBM Granite 4.0（两个变体）、Laguna-XS.2、Hunyuan Hy3、DeepSeek-OCR-2、LLaVA 视频变体等；移除 huihui-ai/grok-2 行。所有新增行均为 模型 ID / 系列名 / 两个勾选列的固定表格结构。
4. 提交组织与收尾：7 个 commit 先功能表后模型表拆分提交，并穿插 "update feature requirements and models" 与 "fix docs review" 两次整合修正，说明过程中有 review 反馈被吸收。

以下片段展示两类表格行的典型结构，均来自本 PR 修改后的文档内容。

```
{/* 支持模型表：每行声明一个 HuggingFace 模型在 Ascend NPU 上可用
    列含义：模型 ID / 系列名 / 两个支持状态勾选列 */}

{/* 本 PR 新增的模型示例：覆盖 MoE 大模型、量化变体与多模态模型 */}
<tr>
  <td>Qwen/Qwen3.8-2.4T-A95B</td>
  <td>Qwen3.8</td>
  <td>✔</td>
  <td>✔</td>
</tr>

<tr>
  <td>sgl-npu/Kimi-K3-W4A8</td>
  <td>Kimi</td>
  <td>✔</td>
  <td>✔</td>
</tr>

<tr>
  <td>deepseek-ai/DeepSeek-OCR-2</td>
  <td>DeepSeek-OCR / OCR-2</td>
  <td>✔</td>
  <td>✔</td>
</tr>
```

```
{/* 支持特性表：每行声明一个 Server 参数在 Ascend NPU 上的支持状态
    列含义：参数名 / 默认值 / 取值类型 / 可用机型（A2/A3） */}

{/* 本 PR 新增的参数行：--retraction-policy 支持 length 与 priority 两种策略 */}
<tr>
  <td>--retraction-policy</td>
  <td>length</td>
  <td>length, priority</td>
  <td>A2, A3</td>
</tr>

{/* 状态从 Planned 升级为 A2/A3：--pp-async-batch-depth 默认值也由 None 修正为
0，减少文档与实现的不一致 */}
<tr>
  <td>--pp-async-batch-depth</td>
  <td>0</td>
  <td>Type: int</td>
  <td>A2, A3</td>
</tr>
```

评论区精华

该 PR 没有实质人工 review 交锋。唯一评论是 sglang-npu-bot 的 /tag-and-rerun-ci，最终由 bot 直接 APPROVED 合并。

> 观察：commit 中的 "fix docs review" 表明存在一次线下 review 修正，但讨论内容未留存在 PR 评论区。

这说明 NPU 文档线目前由专项 bot 自动化维护，速度优先，但缺少公开的 peer review 轨迹。

风险与影响

- 文档与实现一致性风险：--moe-data-parallel-size、--pp-async-batch-depth、--lora-strict-loading 等从 Planned/Special for GPU 直接升级为 A2/A3 支持，PR 未提供验证记录或适配代码引用。若实际覆盖不完整，用户上线后才会发现能力缺失。
- 移除 grok-2 无说明：support_models.mdx 删除该行且 body 未解释，既有用户可能误判为停止支持。
- 新增模型缺验证证据：DeepSeek-OCR-2、Hunyuan Hy3、Granite 4.0 等新增行没有 cookbook、精度或性能数据支撑（PR body 的 Accuracy/Speed Tests 均为 N/A）。
- CI 状态：PR Test 通过但 Extra 运行失败（Run #32469616390），原因未在 PR 内说明，且不影响 bot 合并流程。
- 对用户影响：支持矩阵扩大，直接指导 NPU 部署选型；若表格与实际实现漂移，将增加排障成本。

关联脉络

- PR #35508（NPU A3 Kimi-K3 cookbook）与本 PR 同属 NPU 文档线，本 PR 新增的 sgl-npu/Kimi-K3-W4A8 行正好衔接 Kimi-K3 在 NPU 上的部署指引。
- PR #34237（LFM2 检测器修复）对应的 LFM2 模型支持，在本 PR 模型表中落地为 LFM2 系列行。
- PR #35854（AMD DeepSeek-V4 cookbook）显示各硬件平台文档随适配进展同步刷新的通用模式，本 PR 是 NPU 侧的同类维护。

整体来看，这是 NPU 平台能力扩张在文档侧的周期性同步，价值在于及时更新用户可见的能力边界；但建议后续为 "Planned → Supported" 的状态变更补充验证依据，保持文档长期可信。