

# PR #36940 完整报告

sgl-project/sglang

[NPU] [DOC] update supported features on NPU

合并时间: 2026-08-29 10:52

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36940>

## 执行摘要

本PR更新了NPU平台的文档，包括模型支持列表和功能特性支持状态，属于纯文档维护变更。主要目的是同步 NPU 平台的实际支持范围，避免用户基于过时信息配置。变更由 sglang-npu-bot 审核通过，未涉及代码或测试。

## 功能与动机

PR body 仅说明“update supported features on NPU”，无详细背景。推测是为了反映 NPU 平台对模型和参数的最新支持情况，可能对应实际功能的增删或状态变更。

## 实现拆解

1. 模型支持列表更新：在 support\_models.mdx 中移除了 Falcon-H1-34B-Instruct 等 4 个模型条目，减少误导。
2. 特性支持状态调整：在 support\_features.mdx 中修改了多个参数的支持状态标注，例如将 --lora-strict-loading 的标注从 A2, A3 改为 Special for GPU，更精确地表示其仅适用于 GPU 而不适用于 NPU；类似地，将多个 speculative-ngram 参数标注为 Experimental，表明这些功能在 NPU 上尚处实验阶段。
3. 测试与配置：无测试、配置或部署配套变更。

由于是文档变更，无源码可展示。但核心逻辑在于修正参数支持状态标注，例如：

```
<!-- 示例：原标注为 A2, A3 改为更精确的描述 -->
<td>--lora-strict-loading</td>
<td>False</td>
<td>Special for GPU</td>

<td>--speculative-ngram-match-type</td>
<td>BFS</td>
<td>Experimental。BFS 使用基于最近使用的扩展；PROB 使用基于频率的扩展。</td>
```

## 评论区精华

无实质讨论，仅 bot 自动触发 CI 重跑。

## 风险与影响

- 文档准确性风险：移除的模型是否真的不再支持，以及标注为 Experimental 的参数是否准确，若判断偏差可能误导用户。
- 影响范围：影响 NPU 平台的用户文档浏览体验，对代码无影响。

## 关联脉络

无直接关联历史 PR，但属于 NPU 文档系列的持续更新（例如 PR#36476 更新了 NPU 最佳实践），表明团队在持续完善 NPU 平台文档。