PR #22804 完整报告

sgl-project/sglang

[NPU] Modify the parameter name and optional values, and add the parameter restrictions. Modify some parameters supported type.

合并时间：2026-04-14 21:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22804

执行摘要

本次 PR 更新了 Ascend NPU 支持特性文档，主要调整服务器参数的支持状态、重命名参数并添加限制，以提高文档准确性。变更纯属文档维护，对系统运行无直接影响，但有助于用户正确配置 NPU 平台。

功能与动机

动机源于需要修正 NPU 文档中参数支持状态的不准确描述。PR body 简要说明“修改参数名称、可选值和限制”，具体目标是将部分特性标记为“Planned”（计划中）或“Experimental”（实验性），以反映当前实现状态，避免用户误解。

实现拆解

仅修改文件 docs/platforms/ascend/ascend_npu_support_features.md，更新参数表格：

- 支持状态调整：例如 --swa-full-tokens-ratio 从“A2, A3”改为“Planned”，表示该功能尚未稳定支持。
- 参数重命名：--enable-piecewise-cuda-graph 更名为 --enforce-piecewise-cuda-graph，并更新描述以指定支持模型。
- 限制添加：为某些参数添加类型或使用约束，如将 --enable-dynamic-chunking 标记为“Experimental”。

评论区精华

review 中仅有一条来自 gemini-code-assist[bot] 的评论：

> “The flag --enforce-piecewise-cuda-graph is documented here as supported on A2, A3 platforms. However, in the core implementation, this flag is described as 'Used for testing'... it might be more accurate to mark this feature as Experimental.”

该建议指出文档与代码不一致，但讨论未深入，PR 被合并，可能建议未被采纳或已内部协调。

风险与影响

- 风险：主要风险是文档准确性不足，若支持状态标记错误，可能导致用户错误配置 NPU，引发运行时问题。但无代码变更，故无回归、性能或安全风险。

- 影响：对用户而言，文档更准确，提升使用体验；对系统无影响；对团队，是常规文档维护的一部分。

关联脉络

从近期历史 PR 看，本 PR 是 NPU 文档维护系列的延续，与 PR #22799、#22795、#22793、#22707 等类似，均涉及 Ascend NPU 文档的修复和更新，反映了团队对 NPU 平台文档一致性的持续关注。