PR #22697 完整报告

sgl-project/sglang

[NPU] update npu doc

合并时间：2026-04-13 21:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22697

执行摘要

本 PR 更新了 Ascend NPU 支持特性文档，在编码预填充解耦章节新增了 --enable-adaptive-dispatch-to-encoder 参数，并优化了现有布尔参数的功能描述。这是一个纯文档维护性更新，不影响任何代码逻辑，主要面向使用 Ascend NPU 平台的用户和开发者，帮助他们更清晰地了解相关功能配置。

功能与动机

根据 PR 描述，主要动机是“更新 NPU 文档，添加在 Ascend NPU 上支持的特性参数”。具体来说，需要为编码预填充解耦功能添加新的参数支持说明，使文档保持最新状态。

实现拆解

仅修改了一个文档文件，具体变更如下：

文件路径	变更内容
docs/platforms/ascend/ascend_npu_support_features.md	在“Encode prefill disaggregation”章节的表格中： 1. 新增一行 --enable-adaptive-dispatch-to-encoder 参数 2. 优化 --encoder-only 描述为“bool flag (set to launch an encoder-only server)” 3. 优化 --language-only 描述为“bool flag (set to load weights for the language model only)”

评论区精华

review 中只有一条来自 gemini-code-assist[bot] 的评论：

> “The descriptions for some arguments in this table are quite minimal. To improve user experience and clarity, it would be beneficial to include the more descriptive help texts from server_args.py for the boolean flags. This would save users from having to dig into the source code to understand what each argument does.”

这个建议被采纳，体现在最终的文档修改中——优化了布尔参数的描述文本，使其功能说明更清晰。

风险与影响

风险分析：

- 这是一个纯文档更新，不涉及任何代码、配置或测试变更，因此没有技术风险。
- 唯一的潜在风险是文档描述可能不准确，但考虑到这是对现有功能的补充说明，且修改内容简单明确，风险极低。

影响分析：

- 影响范围仅限于使用 Ascend NPU 平台的用户和开发者。
- 帮助他们更清晰地了解编码预填充解耦功能的参数配置，特别是新增的自适应调度参数。
- 对系统运行、性能、兼容性等无任何影响。

关联脉络

从近期历史 PR 可以看出，这是一个 NPU 文档维护系列的一部分：

1. PR #22700 和 #22698 同样修改了同一个文档文件，分别关注参数约束澄清和默认值修正。
2. PR #22687 也是 NPU 相关的文档修复，虽然文件不同但属于同一技术领域。

这表明团队正在系统性地完善 Ascend NPU 平台的文档，特别是服务器参数和功能支持方面的说明，为 NPU 用户提供更准确、完整的配置指南。