# PR #22698 完整报告

sgl-project/sglang

[Docs] Fix default values and options in Ascend server arguments documentation

合并时间：2026-04-13 21:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22698

## 执行摘要

此 PR 更新了 Ascend NPU 后端服务器参数文档，修正了多处过时的默认值、选项列表和支持状态描述，确保用户获得准确的配置指导。变更仅涉及文档文件，无代码逻辑修改，风险低，影响范围限于 Ascend NPU 用户。

## 功能与动机

文档中存在多处过时或错误信息，导致用户配置困惑。根据 PR body，关键问题包括：--speculative-draft-load-format 的默认值错误（应为 auto 而非 None）、--tool-call-parser 的选项列表未反映当前 Ascend 后端支持的解析器、--disaggregation-decode-enable-offload-kvcache 的支持状态描述不清晰（实际为计划支持而非已支持），以及 HTTP 服务器部分存在重复表格条目。PR 旨在确保用户有正确的配置指导，并明确哪些功能尚在规划中。

## 实现拆解

修改仅涉及一个文件：docs/platforms/ascend/ascend_npu_support_features.md。具体更新如下：

| 部分 | 修改内容 | 影响 |
| --- | --- | --- |
| HTTP Server | 移除重复表格块；将 --grpc-mode 的“Server supported”从“A2, A3”改为“Planned” | 明确 gRPC 模式为计划支持，避免误导 |
| API related | 更新 --tool-call-parser 的“Options”为 llama3、pythonic、qwen、qwen3_coder；微调 --reasoning-parser 格式 | 精简选项列表，匹配当前 Ascend 后端实现 |
| Speculative decoding | 将 --speculative-draft-load-format 的“Defaults”从 None 改为 auto；扩展 --speculative-draft-model-revision 的“Options”描述 | 修正默认值，提供更清晰的版本选项示例 |

| 部分 | 修改内容 | 影响 |
| --- | --- | --- |
| Disaggregation | 将 --disaggregation-decode-enable-offload-kvcache 的“Server supported”从“A2, A3”改为“Planned” | 准确表明 KV 缓存卸载功能尚未支持 |

## 评论区精华

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

> There is a typo in the description: 'dosn't' should be 'doesn't'.

该评论指出 --disaggregation-decode-enable-offload-kvcache 描述中的拼写错误，可能已在提交中修复。sglang-npu-bot 直接批准，无其他技术讨论。

## 风险与影响

- 风险：文档更新可能未完全同步实际代码行为，例如如果 --tool-call-parser 的选项列表或 --speculative-draft-load-format 的默认值与后端实现不一致，可能导致用户配置错误。但基于 PR 动机，这些修正确认了当前状态，风险较低。
- 影响：仅影响 Ascend NPU 后端的用户和开发者，提供更准确的配置文档，减少因文档错误导致的配置困惑或支持请求。不改变系统行为，影响程度低。

## 关联脉络

- 与 PR #21908（Intel GPU 文档更新）类似，同为平台特定文档维护，但针对不同硬件。
- 与 PR #22594（扩散模型量化文档修复）类似，涉及文档同步，但本 PR 为纯文档修正，不涉及代码变更。
- 近期历史 PR 中未见直接关联的 Ascend NPU 功能变更，表明此 PR 为独立的文档清理工作。