

PR #23378 完整报告

sgl-project/sglang

[NPU] offloading docs update

合并时间: 2026-04-22 11:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23378>

执行摘要

本次 PR 更新了 Ascend NPU 平台支持特性的文档，针对 `--offload-mode` 参数的 `sharded_gpu` 选项添加了配置限制说明，明确该模式仅适用于张量并行度 (tp) 为 1 且数据并行度 (dp) 大于 1 的场景。变更仅涉及一行文档修改，风险极低，旨在提高文档准确性以避免用户误配置。

功能与动机

从 PR 标题和 commit 信息推断，本次变更的动机是更新 NPU 相关的卸载文档。具体而言，需要澄清 `--offload-mode` 参数中 `sharded_gpu` 选项的适用配置，确保文档准确反映功能限制，防止用户在错误的并行配置下使用该模式。

实现拆解

- 变更入口：修改 docs/platforms/ascend/ascend_npu_support_features.md 文件，这是 NPU 平台功能支持的主要文档。
- 核心内容更新：在文档表格中，针对 `--offload-mode` 参数的 `sharded_gpu` 选项，从原有描述 `sharded_gpu (DeepSeek only)` 扩展为 `sharded_gpu (DeepSeek only, only support tp=1 dp>1)`，以明确该模式仅支持特定并行配置。
- 无配套改动：本次变更仅涉及文档更新，未修改任何源码、测试、配置或部署文件，因此无需额外步骤。

评论区精华

Review 评论为空，仅有两个由 sglang-npu-bot 执行的自动化批准，表明本次文档更新未引发技术讨论或争议，属于常规维护性变更。

风险与影响

- 技术风险：极低，仅文档更新，不涉及代码逻辑、性能、安全或兼容性变更。潜在风险在于文档说明本身若不准确可能误导用户，但本次变更旨在提高准确性。
- 影响范围：主要影响使用 Ascend NPU 平台并配置 `--offload-mode=sharded_gpu` 的用户，帮助他们避免在 `tp≠1` 或 `dp≤1` 的场景下误用该模式。对系统和团队无实质性影响，仅减少用户支持负担。

关联脉络

- 与近期 PR #23409（环境变量默认值变更）和 #23395（文档重定向更新）类似，都属于文档或配置相关的维护性更新，反映了团队对文档准确性和用户体验的持续关注。
- 在 NPU 平台支持方面，本次更新是功能文档的细化，与仓库中其他 NPU 相关变更（如性能优化或 bugfix）形成补充，确保用户文档与代码实现保持一致。