PR #27035 完整报告

sgl-project/sglang

docs: add DeepSeek V4 FP4 indexer usage

合并时间：2026-06-04 15:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27035

执行摘要

- 一句话：新增 DeepSeek V4 FP4 Indexer 文档
- 推荐动作：文档清晰简洁，可以直接合并。无需精读。

功能与动机

为 DeepSeek V4 FP4 C4 Indexer 实验性功能提供用户文档，该功能用于 SM100 GPU 上 decode-heavy 长上下文场景，旨在降低 indexer 缓存带宽。

实现拆解

1. 更新 Cookbook 使用指南：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中添加 "FP4 Indexer (Experimental)" 章节，说明参数用途、硬件要求（SM100 + DeepGEMM FP4 indexer support）和命令行示例。
2. 更新服务端参数表：在 docs_new/docs/advanced_features/server_arguments.mdx 的配置表格中新增一行，描述 --enable-deepseek-v4-fp4-indexer 的用途、默认值（False）和类型。
3. 移除旧文档中重复变更：根据 review 意见，删除了 docs/advanced_features/server_arguments.md 中的无关改动。

关键文件：

- docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx（模块 文档；类别 other；类型 core-logic）：新增 FP4 Indexer 使用指南和命令示例，是用户主要参考入口。
- docs_new/docs/advanced_features/server_arguments.mdx（模块 文档；类别 other；类型 core-logic）：服务端参数表新增一行，便于用户快速查阅参数含义。

关键符号：未识别

评论区精华

Reviewer Fridge003 指出不应修改旧版文档 docs/advanced_features/server_arguments.md，作者已移除该改动。另外，Fridge003 建议将 Cookbook 中的标题从 "FP4 Indexer" 改为 "FP4 Indexer (Experimental)" 并在代码示例前添加 bash Command 标签，作者均予以采纳。

- 移除旧文档中的变更 (other): 作者移除了该文件的改动。
- Cookbook 中标题和代码块格式修正 (style): 作者已采纳并应用修改。

风险与影响

- 风险：无风险。仅文档变更，不影响任何代码逻辑。
- 影响：对用户而言，提供了新功能的使用说明，有助于用户了解和使用 FP4 Indexer。对系统无影响。影响范围局限于 DeepSeek V4 用户和 CUDA SM100 平台。
- 风险标记：暂无

关联脉络

- PR #26209 [WIP] DeepSeek V4 FP4 indexer: 本 PR 文档对应的功能实现 PR。