PR #22238 完整报告

sgl-project/sglang

[HiSparse]: Add readme docs for HiSparse Feature

合并时间：2026-04-07 15:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22238

执行摘要

- 一句话：新增 HiSparse 分层稀疏注意力功能的使用文档和配置指南。
- 推荐动作：建议文档维护者和使用 HiSparse 功能的工程师阅读此 PR，以了解功能细节和配置方法。对于代码工程师，此 PR 无需精读，但可以作为文档示例参考或了解新功能背景。

功能与动机

从 PR 标题和新增文档内容推断，HiSparse 是一个新功能，需要文档来指导用户使用。文档中说明 HiSparse 用于减少 GPU 内存消耗并提高解码并发性，特别是针对使用 DeepSeek Sparse Attention 架构的模型（如 DeepSeek-V3.2、GLM-5），以支持长期上下文 LLM 推理。

实现拆解

实现方案主要包括：1) 新增 docs/advanced_features/hisparse_guide.md 文件，提供 HiSparse 的完整指南，涵盖简介、设计、服务器参数和配置；2) 修改 docs/basic_usage/deepseek_v32.md 文件，添加对 HiSparse 的引用和简要说明，确保文档一致性。

关键文件：

- docs/advanced_features/hisparse_guide.md（模块 documentation）：新增的 HiSparse 完整指南，包含设计、工作流程、服务器参数和配置说明，是用户使用该功能的核心文档。
- docs/basic_usage/deepseek_v32.md（模块 documentation）：更新 DeepSeek-V3.2 文档以引用 HiSparse 指南，确保文档一致性和用户易用性。

关键符号：未识别

评论区精华

review 评论中，Fridge003 指出文档中模型名称应为 'Deepseek-V3.2' 而不是 'DeepSeek v3'，作者 hzh0425 回复已更新；ShangmingCai 建议未来当支持更多后端和 kvcache 数据类型时，可将此文档添加到 docs/index.rst 中。讨论焦点在于文档准确性和未来维护。

- 模型名称更正 (correctness): 作者 hzh0425 回复 'yes, updated'，表示已修复此错误。
- 未来文档更新建议 (documentation): 无立即行动，作为未来计划，团队可在相关功能扩展时跟进。

风险与影响

- 风险：风险较低，主要涉及文档准确性：如果文档描述错误（如模型名称、配置参数），可能导致用户配置不当或误解功能。具体文件 hisparse_guide.md 中关于技术细节需要确保正确。此外，文档的完整性可能不足，但 ShangmingCai 的评论已指出未来更新点，风险可控。
- 影响：对用户：提供了 HiSparse 功能的使用指南，有助于用户优化长期上下文 LLM 推理的内存和并发性，特别是针对 DeepSeek 稀疏注意力模型。对系统：无直接代码变更，不影响运行时性能。对团队：完善了文档库，提高了功能可发现性和易用性。影响程度：中等，因为文档对功能推广和用户采用至关重要。
- 风险标记：文档准确性风险，未来维护需求

关联脉络

- 暂无明显关联 PR