PR #21716 完整报告

sgl-project/sglang

[Doc] Update GLM-5 instructions in sglang documentation

合并时间：2026-04-05 18:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21716

执行摘要

该 PR 更新了 DeepSeek V3.2 使用文档，将 GLM-5 模型纳入同一指南，因为两者共享 DeepSeek 稀疏注意力（DSA）结构。文档标题改为“DeepSeek V3.2/GLM-5 Usage”，并补充了 GLM-5 的启动命令和配置提示。这是一个低风险的文档维护变更，旨在提升用户体验和文档一致性。

功能与动机

根据 PR 动机清单，需要更新文档以包含 GLM-5 模型的使用说明。GLM-5 模型也应用了 DeepSeek 稀疏注意力（DSA）结构，因此可以与 DeepSeek V3.2 共享大部分使用方式，但推理解析器和工具调用解析器除外。文档更新旨在避免用户重复查阅，并提供统一的配置指导。

实现拆解

修改了单个文件 docs/basic_usage/deepseek_v32.md，关键变更点包括：

1. 标题更新：从“DeepSeek V3.2 Usage”改为“DeepSeek V3.2/GLM-5 Usage”。
2. 内容精简：移除了过时的 Roadmap 链接（原指向 Issue #11060）。
3. GLM-5 集成：
   - 在介绍部分添加说明：“GLM-5 model also applies DSA(Deepseek sparse attention) structure, so can share most of the usage here, except for reasoning parser and tool call parser.”
   - 在启动部分添加命令：“To server GLM-5, just replace the --model argument with zai-org/GLM-5-FP8.”
   - 在配置提示部分补充 GLM-5 注意事项。

这些变更使文档更简洁，并覆盖了 GLM-5 用户的需求。

评论区精华

该 PR 没有技术 review 评论，仅有一个来自 gemini-code-assist[bot] 的 Issue 评论，提示每日配额限制，与 PR 内容无关。因此没有讨论交锋或决策过程。

风险与影响

风险分析：

- 文档准确性风险：需确保 GLM-5 与 DeepSeek V3.2 共享 DSA 结构的表述正确，且差异（如推理解析器）被明确标注。
- 无代码变更，因此无回归、性能、安全或兼容性风险。

影响分析：

- 用户影响：GLM-5 用户现在可以从同一文档获取配置指导，减少学习成本。
- 系统影响：无。
- 团队影响：文档维护更集中，但需注意未来模型差异的及时更新。

关联脉络

从近期历史 PR 看，该 PR 与以下 PR 相关：

1. PR #21405：启用了 DeepSeek V3.2 的 IndexCache 优化，当前 PR 的文档可能隐含了相关配置建议。
2. PR #22140 和 #22108：都涉及 DeepSeek 相关测试或脚本修复，共享 deepseek 标签，反映了团队对 DeepSeek 模型生态的持续投入。

整体上，该 PR 是 DeepSeek 模型文档维护的一部分，旨在保持文档与模型支持同步，符合仓库近期强调的“consistency”趋势（如 PR #22148、#22147 等）。