PR #22179 完整报告

sgl-project/sglang

[Doc] Fix and improve DeepSeek V3.2/GLM-5 documentation

合并时间：2026-04-06 14:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22179

执行摘要

本次 PR 对 DeepSeek V3.2 和 GLM-5 模型的使用文档进行了修正，主要移除了关于 skip-softmax 的错误描述（根据 flashinfer 库约束，该功能不适用于 DSA 稀疏注意力），并修复了多处拼写、语法和术语格式问题。变更仅涉及单个文档文件，不影响代码逻辑，风险极低，但遗留了一个无效的 arXiv 链接占位符待后续处理。

功能与动机

为什么做：根据 PR body 中的说明，作者发现文档中关于 skip-softmax 的描述存在技术错误。引用 flashinfer 库源码（flashinfer/mla.py 第 730-731 行）的约束条件：

```
if skip_softmax_threshold_scale_factor is not None and sparse_mla_top_k != 0:
raise ValueError("skip_softmax is not supported for sparse MLA")
```

这表明 skip_softmax 仅适用于密集注意力（dense attention），而不适用于 DeepSeek Sparse Attention（DSA）的稀疏 MLA。因此需要移除相关描述以避免用户误用。同时，作者还希望改进文档的表述准确性，修复如“server GLM-5”拼写错误等问题。

实现拆解

仅修改了 docs/basic_usage/deepseek_v32.md 文件，变更可分为三类：

1. 技术描述修正：
   - 移除 skip-softmax 相关段落（原文档可能暗示其适用于 DSA）。
   - 统一术语大小写：将“DSA(Deepseek sparse attention)”改为“DSA (DeepSeek Sparse Attention)”。
2. 语言表述优化：
   - 拼写修正：“server GLM-5” → “serve GLM-5”。
   - 语法补充：为“reasoning parser”添加定冠词“the”。
   - 句子结构调整以提升可读性。
3. 格式微调：
   - 调整部分标点和空格使用。

变更总计 11 行新增、12 行删除，均为文本内容调整，无代码逻辑变动。

评论区精华

review 中仅有一条实质性讨论：

> gemini-code-assist[bot]指出："The arXiv link https://arxiv.org/abs/2603.12201 appears to be a placeholder, as it points to a future date and is not a valid arXiv ID format. This could be confusing for readers."

该评论指出文档中一个 arXiv 链接指向未来日期（2603 年），显然是占位符，可能误导读者。但作者和审阅者均未回复此问题，且后续提交未修改该链接，导致问题遗留。Fridge003 作为审阅者直接批准了 PR，并在关联 Issue 中感谢作者发现 skip-softmax 问题。

风险与影响

风险分析：

- 技术风险几乎为零：纯文档变更，不涉及任何代码执行逻辑。
- 移除 skip-softmax 描述有明确依据（flashinfer 库约束），不会引入错误配置。
- 唯一遗留风险是无效 arXiv 链接可能影响文档可信度，但不会导致功能问题。

影响分析：

- 对用户：帮助正确理解 DeepSeek V3.2/GLM-5 的 DSA 特性，避免错误尝试使用 skip-softmax；提升文档可读性。
- 对系统：无任何运行时影响。
- 对团队：文档维护更准确，但遗留链接问题需后续跟进。

关联脉络

从近期历史 PR 看，本 PR 与以下 PR 存在关联：

1. PR #22006（DeepSeek V3 路由方法修复）和 PR #22143（DeepSeek V2 量化格式检测缓存）：
   - 同属 DeepSeek 模型相关改进，但那些 PR 涉及代码 bugfix 和性能优化，而本 PR 是纯文档修正。
   - 反映团队对 DeepSeek 模型生态的持续维护，涵盖代码、性能、文档多方面。
2. 文档维护趋势：
   - 近期多个 PR（如 #22189、#21921、#22111）都包含文档更新，表明团队重视文档与代码同步。
   - 本 PR 延续了这一趋势，针对具体模型（DeepSeek/GLM-5）的技术细节进行校准。

本 PR 虽小，但体现了文档基于依赖库约束及时修正的重要性，避免用户因文档错误而产生配置失误。