PR #23628 完整报告

sgl-project/sglang

[codex] docs: note H200 DeepSeek-V4 checkpoint

合并时间：2026-04-24 15:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23628

执行摘要

在 DeepSeek-V4 部署文档中增加 H200 检查点提示，引导用户使用 sgl-project 下的 SGLang 检查点而非默认 DeepSeek 检查点。变更仅 4 行，无代码逻辑改动。

功能与动机

H200 GPU 不支持 DeepSeek-V4 官方检查点使用的 FP4 MoE 专家，直接使用官方检查点会导致部署失败。本次 PR 通过文档提示避免用户踩坑。

实现拆解

在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中，在第 3.1 节前插入 <Note> 组件：

```
<Note>
For H200 GPU deployments, use the SGLang checkpoint under `sgl-project`, not the default DeepSeek checkpoint.
</Note>
```

未涉及测试、配置或部署配套变更。

评论区精华

AI 审查机器人 gemini-code-assist[bot] 指出：

> 该提示应补充技术原因（Hopper 缺少 FP4 支持），并与第 133 行已有的 H200 备注合并以避免冗余。

该建议未被采纳即合并。

风险与影响

- 风险：极低，仅文档变更。但新增提示与第 133 行现有备注内容重复，若后续不统一可能让用户困惑。
- 影响：仅影响阅读 DeepSeek-V4 部署文档的 H200 用户，帮助他们避免部署失败。

关联脉络

当前 PR 属于 DeepSeek-V4 文档迭代系列的后继更新（#23605 → #23622 → #23617 → #23634 → #23628），持续补充部署细节。建议后续将两个重复的 H200 备注合并，并补充技术原因。