PR #23634 完整报告

sgl-project/sglang

Update pro fp8 checkpoint in DeepSeek V4 cookbook

合并时间：2026-04-24 15:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23634

执行摘要

本 PR 更新了 DeepSeek-V4 部署文档生成器中的 H200 Pro 模型仓库地址，将占位符替换为实际可用的公共仓库，同步修正了注释文字。变更极小，仅涉及一行配置和一行注释，无代码逻辑改动。

功能与动机

之前 DeepSeek-V4-Pro-FP8 权重尚未公开，因此在文档中使用了占位符 "<TO_BE_UPLOADED_DeepSeek-V4-Pro-FP8>"。现在该仓库已公开可用（sgl-project/DeepSeek-V4-Pro-FP8），需要更新文档以消除占位符，让用户能直接复制生成的命令部署。

实现拆解

1. 更新模型仓库 slug：在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 的 HW_SIZE_SPEC 对象中，将 h200lbig 条目的 slug 从 "<TO_BE_UPLOADED_DeepSeek-V4-Pro-FP8>" 改为 "sgl-project/DeepSeek-V4-Pro-FP8"。
2. 更新注释：同步将旁边的注释从 "Flash is public, Pro is still being uploaded" 改为 "repackagings for both variants"，表明两个变体（Flash 和 Pro）均已发布 FP8 权重。

这两个改动确保 H200 Pro 部署命令能生成有效模型路径，无需用户手动替换。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

唯一变更文件，更新了 H200 Pro 的模型仓库 slug 和相关注释。

关键源码片段

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

唯一变更文件，更新了 H200 Pro 的模型仓库 slug 和相关注释。

```
// 在 HW_SIZE_SPEC 中，h200lbig 的 slug 从占位符更新为实际仓库
const HW_SIZE_SPEC = {
  ...
  // sgl-project 为两个变体发布了 FP8 权重重封包
  "h200lsmall": { slug: "sgl-project/DeepSeek-V4-Flash-FP8", tp: 4, multinode: false },
  "h200lbig": { slug: "sgl-project/DeepSeek-V4-Pro-FP8", tp: 16, multinode: true, nnodes: 2 },
};
```

评论区精华

审查者 gemini-code-assist 仅确认“没有进一步反馈”，无实质性讨论。

风险与影响

- 风险：极低。仅字符串替换和注释更新，不影响任何运行时行为。若仓库地址拼写错误，用户手动修正即可。
- 影响：直接影响阅读文档的 H200 Pro 用户，他们现在可获得可用的部署命令；间接影响团队维护成本。

关联脉络

本 PR 与 #23617（Further update Deepseek V4 docs）属于同一功能线，都是 #23605（Add DeepSeek V4 cookbook）的后续文档完善。这三者共同构成了 DeepSeek V4 部署文档的完整发布流程。