PR #23617 完整报告

sgl-project/sglang

Further update Deepseek V4 docs

合并时间：2026-04-24 13:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23617

执行摘要

本 PR 更新了 DeepSeek V4 部署文档中 H200 硬件配置的模型仓库地址，将之前占位符替换为已公开的 FP8 重打包仓库 sgl-project/DeepSeek-V4-Flash-FP8，并修正了相关注释说明。变更非常小（+5/-6），仅涉及一个文件，无技术风险。

功能与动机

让 H200 用户能够直接复制使用正确的 FP8 模型仓库地址，而不是遇到占位符报错。原始文档因等待 Hopper checkpoint 上传而使用了占位符，现在 Flash 模型的 FP8 重打包已公开，需要更新以提供可用命令。

实现拆解

1. 编辑文档代码片段文件docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx：修改了 HW_SIZE_SPEC 中 H200 small 条目的 slug 字段，从占位符字符串 "<TO_BE_UPLOADED_DeepSeek-V4-Flash-hopper>" 改为实际可用的 Hugging Face 仓库 ID"sgl-project/DeepSeek-V4-Flash-FP8"。
2. 更新注释：将原来的注释从说明正在等待上传，改为解释为什么需要 FP8 独立检查点（deepseek-ai 仓库的 FP4 混合权重在 Hopper 上无法运行）以及 sgl-project 已经发布 FP8 重打包，并指明 Flash 版本已公开，Pro 版本仍在等待上传。
3. 未更改其他配置：H200 big 条目仍保留占位符，等待 Pro FP8 重打包上传后更新。
4. 无测试或配置配套改动：本次变更仅为文档源码更新，不涉及测试或部署配置。

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

唯一变更文件，包含 H200 硬件配置的模型仓库地址和注释更新。

关键源码片段

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

唯一变更文件，包含 H200 硬件配置的模型仓库地址和注释更新。

```
// 变更后的 H200 配置片段
const HW_SIZE_SPEC = {
  "b200|small": { slug: "deepseek-ai/DeepSeek-V4-Flash", tp: 4, multinode: false },
  "b200|big": { slug: "deepseek-ai/DeepSeek-V4-Pro", tp: 8, multinode: false },
  "gb300|small": { slug: "deepseek-ai/DeepSeek-V4-Flash", tp: 4, multinode: false },
  "gb300|big": { slug: "deepseek-ai/DeepSeek-V4-Pro", tp: 4, multinode: false },
```

```
  // H200 needs an FP8-only Instruct ckpt (deepseek-ai's Flash/Pro repos ship
  // FP4-mixed weights that Hopper can't run). sgl-project publishes FP8
  // repackagings; Flash is public, Pro is still being uploaded.
  "h200lsmall": { slug: "sgl-project/DeepSeek-V4-Flash-FP8", tp: 4, multinode: false },
  "h200lbig": { slug: "<TO_BE_UPLOADED_DeepSeek-V4-Pro-FP8>", tp: 16, multinode: true,
  nnodes: 2 },
};
// 注释解释了为什么 H200 需要单独的 FP8 检查点，以及现在 Flash FP8 已可用。
```

评论区精华

无讨论内容。

风险与影响

风险极低。变更仅为文档中的字符串和注释更新，不涉及任何运行时逻辑或配置。唯一潜在风险是用户可能误认为 Pro 模型的 FP8 也已可用（但占位符仍在，且注释明确说明了状态），但这属于文档清晰度问题，不构成技术风险。

影响范围小，仅影响查看 DeepSeek V4 部署文档的 H200 用户。正面影响是 H200 Flash 用户现在可以直接复制命令，无需手动替换占位符。

关联脉络

无直接关联的 PR。该文档属于 DeepSeek V4 系列模型的持续部署文档维护的一部分，反映了 sgl-project 开始发布 FP8 重打包模型的生态进展。后续当 Pro 版 FP8 重打包上传后，可快速跟进更新。