PR #23605 完整报告

sgl-project/sglang

Add DeepSeek V4 cookbook

合并时间：2026-04-24 13:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23605

执行摘要

新增 DeepSeek-V4 部署交互指南，包含一个交互式命令生成组件（JSX）和一个详细的使用指南（MDX），帮助用户根据硬件平台、模型变体和部署策略一键生成正确的 sglang serve 命令。同时更新了文档导航和首页入口。变更以文档为主，风险较低。

功能与动机

DeepSeek-V4 模型发布后，用户需要一个清晰、可操作的部署指南。该 PR 通过交互式矩阵，将复杂的部署参数组合（3 种硬件×2 种模型大小×5 种 Recipe×2 个解析器开关）转化为直观的选择界面，并自动生成命令，降低用户误配置风险。

实现拆解

1. 交互式命令生成组件 文件：docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx （+569 行）
- 定义 DeepSeekV4Deployment React 组件，内含 options 配置对象，声明硬件平台（B200/GB300/H200）、模型变体（Flash/Pro）、部署策略（low-latency/balanced/max-throughput/cp/pd-disagg）、推理解析器和工具调用开关。
- 核心函数 generateCommand 根据用户当前选中的值，拼接 CLI 命令。对于尚未在真实检查点上验证的配方（即不在 VERIFIED_RECIPES 集合中的），整个命令块被注释掉，用户复制粘贴后不会意外执行未经验证的配置。
- 暗色模式通过 MutationObserver 监听 <html> 的 class/data-theme/style 变化自动适配。

1. 使用指南文档 文件：docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx （+453 行）
- 以表格形式列出 Flash（284B）和 Pro（1.6T）的参数量、激活参数量及适用场景。
- 说明关键特性：混合注意力（SWA+MLA）、DeepEP 通信、FP4 MoE 专家等。
- “配置建议”小节指导用户使用上一步的交互组件生成命令，并提示并发度和 DeepEP 缓冲区大小的调整。
- 性能基准测试指向独立的生成器页面，避免正文内分散启动命令。

1. 导航注册 文件：docs_new/docs.json （+1 行）
- 在 DeepSeek 分组中插入 cookbook/autoregressive/DeepSeek/DeepSeek-V4，使新页面出现在侧边栏。

1. 首页入口更新 文件：docs_new/cookbook/autoregressive/intro.mdx （+1/-1 行）

- 将 DeepSeek 卡片链接从旧的 V3_2 改为 V4，确保用户从自动回归模型首页直接跳转到最新文档。

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

核心交互组件，定义部署矩阵和命令生成逻辑，是本次变更的技术主体。

关键源码片段

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

核心交互组件，定义部署矩阵和命令生成逻辑，是本次变更的技术主体。

```
// 定义部署矩阵的五个选项组：硬件、模型大小、部署策略、推理解析器、工具调用
const options = {
  hardware: {
    name: "hardware",
    title: "Hardware Platform",
    items: [
      { id: "b200", label: "B200 (FP4)", default: true },
      { id: "gb300", label: "GB300 (FP4)", default: false },
      { id: "h200", label: "H200 (FP8)", default: false },
    ],
  },
  modelSize: {
    name: "modelSize",
    title: "Model Variant",
    items: [
      { id: "small", label: "Flash", default: true, subtitle: "285B" },
      { id: "big", label: "Pro", default: false, subtitle: "1.6T" },
    ],
  },
  // ... recipe, reasoningParser, toolcall 类似定义
};

// 根据当前选择的选项拼接启动命令
export const DeepSeekV4Deployment = () => {
  const [values, setValues] = useState(getInitialState);
  // ...
  const generateCommand = (hardware, modelSize, recipe, parser, toolcall) => {
    // 基于硬件和模型大小确定张量并行度、节点数等参数
    // 并拼接 --model-path 等标志
    // 对于未验证的配方，整块代码被注释掉，用户复制时不会生效
  };
  // 渲染单选按钮和结果命令区域
};
```

评论区精华

无实质 review 讨论。仅有的机器人评论提供了 Mintlify 预览和 API 配额警告，无技术交锋。

风险与影响

- 风险：H200 FP8 检查点尚未公开，生成器会输出 <TO_BE_UPLOADED> 占位符；已验证的配方仅为 B200 small/big 的 low-latency，其他配方均被注释，用户若手动取消注释可能使用未测试的命令。生成器与 sunrise_allinone.py 强耦合，未来模型配置更新需同步修改。
- 影响：正面为主，为用户提供标准化的部署起点，降低 DeepSeek-V4 的上手成本。对现有功能无影响，纯文档添加。

关联脉络

- 23617 （后续 PR）已跟进修复本 PR 中 H200 文档的模型仓库地址。
- 与 #23493、#23545 等 MoE bugfix 无关，但与 DeepSeek 模型家族的文档演进（如 #22774 MUSA 后端支持）共同完善了 SGLang 对 DeepSeek 的全面支持。