PR #25360 完整报告

sgl-project/sglang

[NEW MODEL] Add Ring-2.6-1T cookbook

合并时间：2026-05-15 14:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25360

执行摘要

本 PR 为 InclusionAI 的万亿参数推理模型 Ring-2.6-1T 添加了完整的部署文档与交互式命令生成器。变更包括 MDX 介绍页面、React 交互组件和导航配置，作者在 GB300 x4 和 B200 x8 上完成了烟雾测试。整体为纯文档新增，无后端代码变更，风险极低。

功能与动机

使 SGLang 用户能够轻松部署 Ring-2.6-1T 模型。该模型需要使用 --trust-remote-code 和特定的 TP 与内存配置，文档提供明确指引并允许用户通过交互式组件动态生成部署命令，降低了用户上手成本。

实现拆解

1. 创建模型介绍文档：在 docs_new/cookbook/autoregressive/InclusionAI/Ring-2.6-1T.mdx 中撰写模型介绍、特性、安装步骤、部署命令及配置提示，并引用交互式组件。
2. 创建交互式部署命令生成器：在 docs_new/src/snippets/autoregressive/ring-26-1t-deployment.jsx 中实现 React 组件 Ring261TDeployment，提供硬件平台（GB300 x4 / B200 x8）、Tool Call Parser、Reasoning Parser 的单选按钮，动态组装 sglang serve 启动命令。组件还包含暗色模式自适应。
3. 注册站点导航：在 docs_new/docs.json 的 InclusionAI 分组 pages 数组首部插入新建页面路径 cookbook/autoregressive/InclusionAI/Ring-2.6-1T。
4. 硬件验证：作者分别使用 GB300 x4 和 B200 x8 配置进行烟雾测试，确认 server 健康检查、模型元数据、基本对话、reasoning_effort 参数及工具调用均正常工作，并记录了 GSM8K 性能数据。

以下是交互式命令生成器的核心数据结构与命令生成逻辑（完整组件参见源文件）：

```
export const Ring261TDeployment = () => {
  const options = {
    hardware: { name: 'hardware', title: 'Hardware Platform', items: [{ id: 'gb300', label:
    'GB300 x4', default: true }, { id: 'b200', label: 'B200 x8', default: false }] },
    toolcall: { name: 'toolcall', title: 'Tool Call Parser', items: [{ id: 'enabled', label: 'Enabled',
    default: true }, { id: 'disabled', label: 'Disabled', default: false }] },
    reasoning: { name: 'reasoning', title: 'Reasoning Parser', items: [{ id: 'enabled', label:
    'Enabled', default: true }, { id: 'disabled', label: 'Disabled', default: false }] },
  };
```

```
  const modelConfigs = {
    gb300: { tp: 4, memFraction: '0.95' },
    b200: { tp: 8, memFraction: '0.8' },
  };

  const getInitialState = () => {
    const initialState = {};
    Object.entries(options).forEach(([key, option]) => {
      const defaultItem = option.items.find((item) => item.default);
      initialState[key] = defaultItem ? defaultItem.id : option.items[0].id;
    });
    return initialState;
  };

  const [values, setValues] = useState(getInitialState());
  const [isDark, setIsDark] = useState(false);

  useEffect(() => {
    const checkDarkMode = () => {
      const html = document.documentElement;
      setIsDark(html.classList.contains('dark') || html.getAttribute('data-theme') === 'dark');
    };
    checkDarkMode();
    const observer = new MutationObserver(checkDarkMode);
    observer.observe(document.documentElement, { attributes: true, attributeFilter: ['class',
    'data-theme'] });
    return () => observer.disconnect();
  }, []);

  const generateCommand = () => {
    const { hardware, toolcall, reasoning } = values;
    const { tp, memFraction } = modelConfigs[hardware];
    let cmd = `sglang serve --model-path inclusionAI/Ring-2.6-1T --tp-size ${tp} --trust-remote-
    code --host 0.0.0.0 --port ${PORT} --mem-fraction-static ${memFraction}`;
    if (toolcall === 'enabled') cmd += ' --tool-call-parser glm';
    if (reasoning === 'enabled') cmd += ' --reasoning-parser deepseek-r1';
    return cmd;
  };
  // 渲染部分（略）
};
```

评论区精华

本次 PR 无 Review 讨论。作者在关联 Issue 评论中提交了以下验证结果：

- B200 x8 烟雾测试: /health 返回 200，/v1/models 返回 max_model_len: 131072，基本对话、reasoning_effort: "high"、chat_template_kwargs: {"reasoning_effort": "xhigh"} 及工具调用均通过。

- GB300 x4 烟雾测试: 同样通过健康检查，GSM8K 得分 0.990，输出吞吐量 621.469 token/s。

风险与影响

- 风险: 纯文档变更，无后端影响。交互组件在主流浏览器上工作正常，但若文档站点主题系统变更可能需要适配。
- 影响: 为 Ring-2.6-1T 用户提供官方部署指南，降低配置门槛；团队需维护一份文档。

关联脉络

本 PR 与 #25369（Add hicache feature in dsv4 cookbook）同属 cookbook 文档添加类型，均涉及 docs_new 目录下的 MDX 文件、交互组件和导航注册。这表明 SGLang 正在持续丰富模型部署文档体系，通过组件化交互式文档提升用户体验。