PR #24730 完整报告

sgl-project/sglang

[Cookbook]: add Laguna-XS.2 (Poolside)

合并时间：2026-05-12 23:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24730

1. 执行摘要

本 PR 为 Poolside 的 Laguna-XS.2 混合 MoE 模型新增了详细的部署 cookbook 页面，并附带一个交互式启动命令生成器 React 组件。同时更新了文档侧边栏和首页博客卡片。变更集中在 docs_new 目录，不涉及运行时代码，风险低，为即将合并的模型支持 PR（#24204）提供配套文档。

2. 功能与动机

Laguna-XS.2 是 Poolside 开源的 33.4B 参数混合 SWA+MoE 模型，专为 agentic coding 设计。在 SGLang 中通过 PR #24204 获得原生支持后，用户需要一个清晰的部署指南来配置带有推理 / 工具调用解析器的服务实例。本 cookbook 以 Gemma4 和 MiMo-V2.5 的文档风格为模板，提供硬件（H200/B200）、量化（BF16/FP8/NVFP4）等多维度选项，并通过交互式组件降低命令行参数记忆负担。

3. 实现拆解

3.1 交互式部署生成器（新增）

文件 docs_new/src/snippets/autoregressive/laguna-xs2-deployment.jsx 是一个 React 无状态组件，通过 useState 管理硬件、量化、推理 / 工具调用解析器和 DP Attention 的选择状态。当用户改变选项时，handleRadioChange 自动更新状态并处理选项间的依赖关系（如 NVFP4 仅在 B200 下可用）。最终的 generateCommand 函数根据当前选项拼接成完整的 sglang serve 命令，支持错误提示（如非法组合）。

3.2 Cookbook 文档页面（新增）

docs_new/cookbook/autoregressive/Poolside/Laguna-XS.2.mdx 包含六个章节：

- 模型介绍（架构、量化版本、许可证）
- SGLang 安装（nightly wheel + Docker，引用交互式组件）
- 启动命令（基础 TP=8 示例，可选 DP Attention）
- 推理解析器（--reasoning-parser poolside_v1）
- 工具调用解析器（--tool-call-parser poolside_v1）
- 基准测试参考（链接到 HF 模型卡）。

3.3 侧边栏与首页配置

- docs_new/docs.json：在 Autoregressive Models 分组下新增 "Poolside" 条目，指向新页面。
- docs_new/index.mdx：替换首页的 LMSYS 博客卡片为 P2P weight transfer 和更新后的 DeepSeek-V4 文章，保持首页新闻时效性。

3.4 提交历史与迭代

共17次提交，包括多次往返修改：初期包含--chat-template-kwargs错误参数（后被删除）、域名修正（.ai → .io）、章节结构调整（安装部分简化为 Gemma4 风格）、以及最终移除非服务器端参数。体现了 cookbook 文档从编写到审查的迭代过程。

docs_new/src/snippets/autoregressive/laguna-xs2-deployment.jsx

核心交互式部署配置生成器，通过选项选择实时生成启动命令，是 cookbook 页面的核心交互部分。

```
export const LagunaXS2Deployment = () => {
  // 选项配置：硬件、量化、推理解析器等
  const options = {
    hardware: {
      name: 'hardware',
      title: 'Hardware Platform',
      items: [
        { id: 'h200', label: 'H200', default: true },
        { id: 'b200', label: 'B200/GB200', default: false }
      ]
    },
    quantization: {
      name: 'quantization',
      title: 'Quantization',
      items: [
        { id: 'bf16', label: 'BF16', default: true },
        { id: 'fp8', label: 'FP8', default: false },
        { id: 'nvfp4', label: 'NVFP4', default: false } // Blackwell-only
      ]
    },
    // reasoning, toolcall, dpAttention 等 ...
  };

  // 量化到模型 ID 的映射
  const modelByQuant = {
    bf16: 'poolside/Laguna-XS.2',
    fp8: 'poolside/Laguna-XS.2-FP8',
    nvfp4: 'poolside/Laguna-XS.2-NVFP4'
  };

  // 生成最终启动命令的核心逻辑
  const generateCommand = () => {
    const { hardware, quantization, reasoning, toolcall, dpAttention } = values;
```

```
    // 检查非法组合：NVFP4 仅支持 B200
    if (hardware === 'h200' && quantization === 'nvfp4') {
      return '# Error: NVFP4 is Blackwell-only. Select B200, or pick BF16/FP8 for H200.';
    }

    const modelId = modelByQuant[quantization];
    if (!modelId) return `# Error: Unknown quantization: ${quantization}`;

    const tp = 8; // 固定 TP=8
    const lines = [
      'sglang serve \\',
      `  --model-path ${modelId} \\`,
      `  --tp ${tp}`
    ];

    if (dpAttention === 'enabled') {
      lines[lines.length - 1] += ' \\';
      lines.push(`  --dp ${tp} \\`);
      lines.push('  --enable-dp-attention');
    }

    if (reasoning === 'enabled') {
      lines[lines.length - 1] += ' \\';
      lines.push('  --reasoning-parser poolside_v1');
    }

    if (toolcall === 'enabled') {
      lines[lines.length - 1] += ' \\';
      lines.push('  --tool-call-parser poolside_v1');
    }

    lines.push('  --host 0.0.0.0');
    return lines.join('\n');
  };
  // ... 其余组件逻辑（状态管理、暗色模式监听等）
```

5. 评论区精华

- JustinTong0323：--chat-template-kwargs 不是服务器端参数，仅用于 API 请求。
  - Jiminator：已修复，并对所有启动参数做了全面审查。
- JustinTong0323：Poolside 博客链接应该是 .io 而不是 .ai。
  - Jiminator：确认 .ai 会重定向，但已按建议修正。

两个 review 问题均已关闭，PR 获得批准。

6. 风险与影响

- 风险：文档中的安装命令依赖尚未发布的 nightly wheel 和 Docker 标签，若发布流程延迟可能导致用户安装失败；交互式生成器仅在作者的 H200 环境中测试，其他硬件组合未充分

验证。

- 影响：降低用户部署 Laguna-XS.2 的门槛；为后续类似复杂模型的 cookbook 提供了可复用的交互式组件模式；首页博客卡片更新确保了 LMSYS 新闻时效性。

7. 关联脉络

本 PR 是 PR #24204（Laguna-XS.2 模型原生支持）的配套文档，两个 PR 协同为用户提供从代码到部署的完整体验。此外，PR 的文档风格参考了近期合并的 Gemma4（#24305）和 MiMo-V2.5（#24612）cookbook，保持了文档体系的一致性。

跨 PR 演进趋势

- SGLang 团队近期加强了第三方模型的文档建设，尤其是混合架构（SWA+MoE）和 agentic 场景的部署指南。
- 交互式组件（如 LagunaXS2Deployment）从早期 cookbook 的静态命令示例演进而来，提升了用户交互体验。