PR #23851 完整报告

sgl-project/sglang

[Docs] add cookbook for MiMo-V2.5 family

合并时间：2026-04-28 01:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23851

执行摘要

该 PR 为 XiaomiMiMo 的 MiMo-V2.5 系列模型（包括 Pro 1.02T 和 V2.5 310B 多模态）创建了完整的 cookbook 文档，核心亮点是一个交互式部署命令生成器（JSX 组件），用户可通过选择模型变体、硬件平台和优化选项自动获取准确的 sglang 启动命令。PR 还包含模型介绍、推理示例和基准数据（GSM8K 98.0%，MMMU 59.3%）。

功能与动机

为最新发布的 MiMo-V2.5 系列提供一站式部署指南，降低用户上手门槛。PR 标题和 body 明确表示“Adds a unified cookbook for MiMo-V2.5 family”，涵盖 Pro 和 base 两种变体，支持多种硬件（H200、H100、B200、GB300）和可选的推理优化（EAGLE MTP、DeepEP、DP Attention 等）。

实现拆解

1. 交互式部署生成器：mimo-v25-deployment.jsx 通过 React 状态管理选项，根据模型变体和硬件计算约束，动态生成完整命令（含多节点和优化标志）。
2. Cookbook 主文档：MiMo-V2.5.mdx 包含模型规格表格、特性列表、部署步骤、推理示例（文本思考、图像 / 视频 / 音频、工具调用）及基准（速度与精度）。
3. 导航配置：docs.json 在 Xiaomi 分组中添加新页面引用，确保侧边栏正确显示。
4. 入口更新：intro.mdx 的 Xiaomi 卡片链接从旧文档指向新 cookbook，并调整顺序。

docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx

新增的交互式部署命令生成器，是 PR 核心创新：通过 React 组件动态生成 sglang 启动命令，涵盖 Pro/V2.5、四种硬件、多种优化选项，并内置约束逻辑（如 DeepEP 仅在 Hopper 可用）。

```
// ================================================
// MiMo-V2.5 部署生成器组件
// ================================================
export const MiMoV25Deployment = () => {
  // 模型变体 × 硬件 → slug, tp, 是否多节点 , 是否 Blackwell
  // 根据此矩阵生成最终命令行
  const HW_VARIANT_SPEC = {
    "proIh200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2,
    blackwell: false },
    "proIh100": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2,
```

```
    blackwell: false },
    "prolb200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: false, blackwell: true },
    "prolgb300": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: true, nnodes: 2,
    blackwell: true },
    "baselh200": { slug: "XiaomiMiMo/MiMo-V2.5", tp: 8, multinode: false, blackwell: false, dp: 2 }
    ,
    "baselh100": { slug: "XiaomiMiMo/MiMo-V2.5", tp: 8, multinode: false, blackwell: false, dp: 2 }
    ,
    "baselb200": { slug: "XiaomiMiMo/MiMo-V2.5", tp: 4, multinode: false, blackwell: true, dp: 1 },
    "baselgb300": { slug: "XiaomiMiMo/MiMo-V2.5", tp: 4, multinode: false, blackwell: true, dp: 1 }
    ,
  };

  // 生成多节点额外标志
  const multiNodeFlags = (nnodes) => [
    `  --nnodes ${nnodes}`,
    `  --node-rank <node-rank>`,
    `  --dist-init-addr <node0-ip>:20000`,
  ];

  // 根据约束条件启用 / 禁用选项
  const computeConstraints = (variant, hardware) => {
    // 例如：Blackwell 禁掉 DeepEP
    const isBlackwell = hardware === "b200" || hardware === "gb300";
    const isPro = variant === "pro";
    return {
      deepepDisabled: isBlackwell, // Blackwell 使用 flashinfer_trtllm
      eagleMtpDisabled: !isPro, // MTP 仅 Pro 支持
      // ... 其他约束
    };
  };

  // ... 渲染逻辑：根据选中的 variant/hardware 组合生成命令行字符串
  // 并在 UI 中显示生成的命令
}
```

评论区精华

PR 获得 wisclmy0611 的快速批准，无额外评论或争议。

风险与影响

风险较低：JSX 组件逻辑可能随后端参数更新而过时，但作者已本地验证；MMMU 基准曾因提取 bug 临时移除（后修复）。影响集中在文档层面，无核心代码变更。

关联脉络

与 PR #23808（MiMo-V2.5-Pro day-zero support）和 PR #23811（MiMo-V2.5 day-zero support）紧密相关，该 cookbook 基于它们的部署配置和基准数据构建。