PR #28060 完整报告

sgl-project/sglang

docs

合并时间：2026-06-12 21:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28060

执行摘要

本 PR 为 MiniMax-M3 模型新增完整的 cookbook 文档，包含 config-driven 的部署矩阵和跨硬件平台性能基准，同时更新导航和简介页面。作为 PR #27944 引入的 M3 模型支持的配套文档，该变更使用户能够快速部署和评估 M3 模型。

功能与动机

MiniMax-M3 是 SGLang 近期支持的新模型（PR #27944），提供了一个 config-driven 的文档模式，用户可根据硬件选择部署参数。此文档的动机是填补官方文档空白，提供安装步骤、配置选项和性能基准，降低用户落地门槛。

实现拆解

1. 核心配置文件 minimax-m3.jsx（新增）：定义模型名称、支持硬件、量化方式（MXFP8/BF16）、Docker 镜像映射、benchmark 命令和 playground 功能旋钮。该文件被 _deployment.jsx 消费，生成交互式部署面板。
2. 基准数据文件 minimax-m3-benchmarks.jsx（新增）：为 B200、H200、B300、GB300、MI355X 等平台提供 TTFT、TPOT 和吞吐量数据，以及 GSM8K 精度数据。所有数据注释了测量条件和版本。
3. 文档页面 MiniMax-M3.mdx（新增）：包含模型简介、安装步骤（Python/Docker）、配置驱动的部署面板和 playground 功能演示。页面顶部标注 "NEW" 标签。
4. 导航配置更新 docs.json（修改）：在 MiniMax 分组中添加 M3 页面，并调整排序为组内首位。
5. 简介页面更新 intro.mdx（修改）：将 MiniMax 卡片链接从 M2.5 改为 M3，使 M3 成为 MiniMax 系列的默认入口。
6. 旧版更新 MiniMax-M2.7.mdx（修改）：移除 tag: NEW 标记，因为 M3 已取代其最新模型地位。

minimax-m3.jsx 核心配置（节选）

```
export const config = {
  modelName: "MiniMax-M3",
  supportedHardware: ["b200", "b300", "gb200", "gb300", "mi300x", "mi325x", "mi350x",
  "mi355x", "h200"],
  quantizations: [
    { id: "mxfp8", label: "MXFP8" },
```

```
    { id: "bf16", label: "BF16" },
  ],
  modelNames: {
    "default|mxfp8": "MiniMaxAI/MiniMax-M3-MXFP8",
    "default|bf16": "MiniMaxAI/MiniMax-M3",
  },
  dockerImages: {
    b200: "lmsysorg/sglang:dev-minimax-m3",
    b300: "lmsysorg/sglang:dev-cu13-minimax-m3",
    h200: "lmsysorg/sglang:dev-cu12-minimax-m3",
    mi355x: "lmsysorg/sglang:<rocm-tag>-rocm720-mi35x",
  },
  // 完整配置见源码
};
```

minimax-m3-benchmarks.jsx 基准数据（节选）

```
export const benchmarks = [
  {
    match: { hw: "b200", variant: "default", quant: "mxfp8", strategy: "balanced", nodes: "single" },
    speed: [{ workload: { isl: 2048, osl: 256, max_concurrency: 64, num_prompts: 128 },
              ttft_ms: 2410, tpot_ms: 148.4, tokens_per_sec_per_gpu: 124 }],
    accuracy: { gsm8k_pct: 94.4 },
  },
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    speed: [{ workload: { isl: 2048, osl: 256, max_concurrency: 64, num_prompts: 128 },
              ttft_ms: 1068, tpot_ms: 78.0, tokens_per_sec_per_gpu: 105 }],
    accuracy: { gsm8k_pct: 97.0 },
  },
];
```

MiniMax-M3.mdx 文档导入与部署组件使用

```
import { Deployment } from "/src/snippets/_deployment.jsx";
import { config } from "/src/snippets/configs/MiniMaxAI/minimax-m3.jsx";
import { benchmarks } from "/src/snippets/configs/MiniMaxAI/minimax-m3-benchmarks.jsx";

<Deployment config={config} benchmarks={benchmarks} />
```

评论区精华

该 PR 无新增 review 评论，由维护者直接合并。

风险与影响

- 风险：基准测试数据可能因测试条件差异不准确，Docker 镜像标签可能随构建更新失效。但均为文档层面，不影响系统运行时。
- 影响：用户获得清晰的部署指导，团队减少重复咨询。影响范围为文档用户，包括 NVIDIA 和 AMD GPU 用户。

关联脉络

- 关联 PR #27944：MiniMax-M3 模型支持由该 PR 引入，本文档是其后继配套文档。基准数据表中明确标注测试版本为 PR #27944。
- 与 PR #28061（docs）类似，本 PR 沿用了 SGLang 文档的 config-driven 部署面板模式。