

PR #28061 完整报告

sgl-project/sglang

docs

合并时间: 2026-06-12 21:19

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28061>

执行摘要

本次变更为 MiniMax-M3 模型添加了一整套 cookbook 文档，包括部署指南、交互式配置面板和跨平台性能基准。通过新增一个页面、两个 JSX 配置文件并更新导航链接，用户可一站式获取在 Blackwell、Hopper、AMD Instinct 等硬件上的启动命令和预期性能。这是一次轻量级、无运行时的文档补充。

功能与动机

MiniMax-M3 是一个约 428B 参数的多模态 MoE 推理模型，支持 1M 上下文和 MiniMax Sparse Attention (MSA)。此前社区虽已通过 PR #27944 添加了模型支持，但缺少配套的官方部署文档。本 PR 的目标是提供标准化的部署流程，帮助用户快速上手并了解不同硬件上的性能表现。

实现拆解

- 主文档页 ([MiniMax-M3.mdx](#)): 采用 SGLang 文档的标准 cookbook 结构，包含安装步骤 (Python 源码安装 / Docker)、硬件选择器、命令生成面板 (基于 `_deployment.jsx`) 和 Playground 实验区。页面通过 `import` 引入部署配置和基准数据。
- 部署配置文件 ([minimax-m3.jsx](#)): 导出 `config` 对象，定义 `supportedHardware` (9 种硬件)、`quantizations` (MXFP8/BF16)、`dockerImages` (按硬件映射)、`benchmarkCommands` (`speed + accuracy` 模板) 以及 `playgroundFeatures` (TP、CP、DP-Attention、MoE 后端等)。该文件被 `_deployment.jsx` 组件消费以生成交互式命令。
- 基准数据文件 ([minimax-m3-benchmarks.jsx](#)): 导出 `benchmarks` 数组，每条记录包含 `match` 条件 (`hw/variant/quant/strategy/nodes`)、`sglang_version`、`speed` (TTFT/TPOT/tok-s) 和 `accuracy` (GSM8K 百分比)。当前包含 B200、H200、B300、GB300、MI355X 等平台的数据，部分为 `stub` 或待更新。
- 导航注册 ([docs.json](#)): 在 MiniMax 分组的 `pages` 数组头部插入 `"cookbook/autoregressive/MiniMax/MiniMax-M3"`，确保页面在侧边栏中显示。
- 入口卡片调整 ([intro.mdx](#) 和 [MiniMax-M2.7.mdx](#)): 将概览页的 MiniMax 卡片 `href` 从 M2.5 改为 M3；移除 M2.7 页面的 NEW 标签，避免误导。

[docs_new/cookbook/autoregressive/MiniMax/MiniMax-M3.mdx](#)

核心文档页面，完整介绍 MiniMax-M3 的部署步骤和特性。

```
// MiniMax-M3.mdx - 主文档页关键导入与组件使用
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";

// 部署面板：根据用户选择的硬件、量化、策略等生成启动命令
<Deployment config={config} benchmarks={benchmarks} />

// Playground：允许用户调整更多参数（TP、CP、MoE 后端等）
// 其配置来自 config.playgroundFeatures
```

docs_new/src/snippets/configs/MiniMaxAI/minimax-m3.jsx

部署配置的核心文件，定义模型名称、硬件支持、量化、Docker 镜像映射、基准命令和 Playground 参数。

```
// minimax-m3.jsx - MiniMax-M3 部署配置
// 被 _deployment.jsx 和 _playground.jsx 使用
export const config = {
  modelName: "MiniMax-M3",
  // 支持硬件列表（已验证或推断支持）
  supportedHardware: [
    "b200", "b300", "gb200", "gb300", // NVIDIA Blackwell
    "mi300x", "mi325x", "mi350x", "mi355x", // AMD Instinct
    "h200", // NVIDIA Hopper
  ],
  variants: [{ id: "default", label: "Default" }],
  quantizations: [
    { id: "mxfp8", label: "MXFP8" }, // 原生 MXFP8, Blackwell 与 MI355X 支持
    { id: "bf16", label: "BF16" }, // BF16 回退, Hopper 使用
  ],
  // 硬件到 Docker 镜像的映射
  dockerImages: {
    b200: "lmsysorg/sglang:dev-minimax-m3",
    b300: "lmsysorg/sglang:dev-cu13-minimax-m3",
    gb200: "lmsysorg/sglang:dev-cu13-minimax-m3",
    gb300: "lmsysorg/sglang:dev-cu13-minimax-m3",
    h200: "lmsysorg/sglang:dev-cu12-minimax-m3",
    mi300x: "lmsysorg/sglang:<rocm-tag>-rocm700-mi30x",
    mi355x: "lmsysorg/sglang:<rocm-tag>-rocm720-mi35x",
  },
  // Playground 中可调整的 MoE 并行选项
  playgroundFeatures: {
    moe: {
      backend: {
        options: [
          { id: null, label: "Inherited" },
          { id: "deepen", label: "DeepEP", flags: ["--moe-a2a-backend deepen"] },
        ],
      },
    },
    ep: { label: "EP", values: [null, 2, 4, 8] },
  },
}
```

```

    attention: { knobs: [ { id: "tp", label: "TP", values: [null, 1, 2, 4, 8] } ] },
  },
// 基准命令模板 (bench_serving 和 sgl-eval)
benchmarkCommands: {
  speed: `python3 -m sglang.bench_serving --backend sglang --host {{CURL_HOST}} --port
{{CURL_PORT}} --model {{MODEL_NAME}} --dataset-name {{DATASET}} --random-input-len
{{ISL}} --random-output-len {{OSL}} --num-prompts {{NUM_PROMPTS}} --max-concurrency
{{MAX_CONCURRENCY}}`,
  accuracy: {
    gsm8k_pct: `pip install git+https://github.com/sgl-project/sgl-eval\nsgl-eval run gsm8k --
base-url http://{{CURL_HOST}}:{{CURL_PORT}}/v1 --model {{MODEL_NAME}} --temperature 1.
0 --top-p 0.95 --thinking`,
  },
},
};

```

docs_new/src/snippets/configs/MiniMaxAI/minimax-m3-benchmarks.jsx

提供各硬件平台下的性能基准数据和 GSM8K 准确性结果。

```

// minimax-m3-benchmarks.jsx - MiniMax-M3 各平台性能基准数据
// 每个对象通过 match 字段与部署配置单元格对应
export const benchmarks = [
  {
    // Blackwell B200, MXFP8, tp4, MSA 路径
    match: { hw: "b200", variant: "default", quant: "mxfp8", strategy: "balanced", nodes: "single" },
    sglang_version: "PR #27944",
    speed: [{ workload: { dataset: "random", isl: 2048, osl: 256, max_concurrency: 64, num_
prompts: 128 },
      tft_ms: 2410, tpot_ms: 148.4, tokens_per_sec_per_gpu: 124 }],
    accuracy: { gsm8k_pct: 94.4 }, // sgl-eval --thinking, 全量 1319 题, 推荐采样参数
  },
  {
    // Hopper H200, BF16, tp8, 内置 Triton 稀疏路径
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    sglang_version: "PR #27944",
    speed: [{ workload: { dataset: "random", isl: 2048, osl: 256, max_concurrency: 64, num_
prompts: 128 },
      tft_ms: 1068, tpot_ms: 78.0, tokens_per_sec_per_gpu: 105 }],
    accuracy: { gsm8k_pct: 97.0 }, // sgl-eval --thinking, 全量 1319 题
  },
  // ... 其他平台 (B300/GB300/MI355X) 数据类似
];

```

评论区精华

无讨论记录。合并者 JustinTong0323 直接批准, 审核过程零评论。

风险与影响

- 风险：低。主要为文档路径正确性和基准数据准确性。Docker 镜像标签中的 `<rocm-tag>` 占位符需要用户自行替换为实际标签，可能增加新手困惑。
- 影响：用户端获得官方部署指南；团队端文档结构更完整；无代码运行时影响。

关联脉络

本 PR 是 PR #27944（添加 MiniMax-M3 模型支持）的配套文档。两者结合为用户提供了从模型加载到性能基准的端到端体验。此外，此 PR 遵循了 SGLang 文档体系中的 cookbook 模板模式，与已有 MiniMax-M2、M2.5 等页面结构一致，便于后续模型扩展。