

# PR #31360 完整报告

sgl-project/sglang

Add Inkling cookbook

合并时间: 2026-07-16 02:24

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31360>

## 执行摘要

此 PR 为 Thinking Machines 的 Inkling 模型添加了完整的 cookbook 文档，用户可通过交互式命令面板快速获取部署命令。变更仅涉及文档站点，无代码逻辑改动。

## 功能与动机

为用户提供 Inkling 模型的官方部署指南，包括硬件支持、变体选择、量化策略、基准测试以及启动命令。该模型是一个 975B MoE 模型，支持 1M token 上下文。目前部分硬件平台和基准测试数据尚未完成验证，文档已明确标注。

## 实现拆解

1. 配置片段 (inkling.jsx) : 定义模型元数据，包括支持的硬件、变体、量化策略、模型名称、Docker 镜像等。该文件是 Deployment 组件的数据源。
2. 基准测试数据 (inkling-benchmarks.jsx) : 列出所有硬件 / 配置组合的基准测试条目，暂缺实测数据，仅定义匹配规则。
3. Cookbook 页面 (Inkling.mdx) : 使用 Deployment 组件集成配置和基准测试，生成交互式命令面板。同时提供手动安装、Docker 启动等补充说明。
4. 导航更新 (docs.json) : 在“自回归模型”菜单新增“Thinking Machines”分组。
5. 首页卡片 (intro.mdx) : 添加指向 Inkling cookbook 的卡片。
6. 品牌徽标 (thinkingmachines.png) : 卡片所需的图片。

[docs\\_new/src/snippets/configs/thinkingmachines/inkling-benchmarks.jsx](#)

基准测试数据定义，列出了所有硬件 / 配置组合的匹配条目。

```
// 每个条目对应一个单元格的 `match` 元组。准确度来自 config.defaultAccuracy
// (模型级别，在推理努力度最大时测量)。速度待定 — 一旦在每个单元格上运行了 bench_serving,
// 就填入 tokens_per_sec_per_gpu / ttft_ms / tpot_ms,
// 并将每个条目的 `sglang_version` 设置为可复现的锚点 (发布版本 / commit / PR)。
```

```
export const benchmarks = [
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
  { match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
  { match: { hw: "gb200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
```

```

{ match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes:
"single" } },
{ match: { hw: "h200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes:
"single" } },
{ match: { hw: "mi350x", variant: "default", quant: "bf16", strategy: "balanced", nodes:
"single" } },
{ match: { hw: "mi355x", variant: "default", quant: "bf16", strategy: "balanced", nodes:
"single" } },
{ match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
{ match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
{ match: { hw: "gb200", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
{ match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
{ match: { hw: "h200", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
{ match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "long_context", nodes:
"single" } },
{ match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "long_context", nodes:
"single" } },
{ match: { hw: "gb200", variant: "default", quant: "nvfp4", strategy: "long_context", nodes:
"single" } },
{ match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "long_context", nodes:
"single" } },
{ match: { hw: "gb300", variant: "default", quant: "bf16", strategy: "balanced", nodes: "multi-
2" } },
{ match: { hw: "gb300", variant: "default", quant: "bf16", strategy: "mtp", nodes: "multi-2" } },
{ match: { hw: "b300", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single"
} },
{ match: { hw: "b300", variant: "default", quant: "bf16", strategy: "mtp", nodes: "single" } },
{ match: { hw: "b200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "multi-
2" } },
{ match: { hw: "b200", variant: "default", quant: "bf16", strategy: "mtp", nodes: "multi-2" } },
{ match: { hw: "b200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" }
},
{ match: { hw: "b300", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" }
},
{ match: { hw: "gb200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" }
},
{ match: { hw: "gb300", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" }
},
{ match: { hw: "h200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" }
},
{ match: { hw: "gb300", variant: "lora", quant: "bf16", strategy: "balanced", nodes: "multi-2" }
},
{ match: { hw: "h200", variant: "lora", quant: "bf16", strategy: "balanced", nodes: "single" } },
];

```

## 评论区精华

无实质 review 讨论。仅有一条来自 gemini-code-assist[bot] 的自动配额警告。

## 风险与影响

- 风险 \*：文档内容（如镜像标签、命令参数）可能随版本过期，但作者已明确标注“未验证”和“待定”状态，风险可控。无代码变更，不存在回归或安全问题。
- 影响 \*：仅影响 sglang 文档站点，为用户提供 Inkling 模型的部署指南。无运行时影响。

## 关联脉络

此 PR 是独立的 cookbook 添加，与近期其他 cookbook（如 DeepSeek-V4 #30651）遵循相同的文档模式。未来可基于此模板添加更多模型。