

PR #32951 完整报告

sgl-project/sglang

Add Inkling-Small cookbook

合并时间: 2026-07-31 01:59

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32951>

执行摘要

本 PR 为 Thinking Machines 的 Inkling-Small 模型添加了完整的部署 cookbook，包括交互式命令生成面板、基准测试精度数据与跨硬件部署策略。同时更新了现有 Inkling 模型的 Docker 镜像方案并增强部署组件的投机解码提示通用性。

功能与动机

Inkling-Small 是 Thinking Machines 新推出的模型，尚无官方部署指南。本 PR 填补了该空白，使用户能够在不同硬件上快速部署模型。

实现拆解

1. 新增 Inkling-Small 配置数据文件 `inkling-small.jsx`: 导出 `config` 对象，包含模型名称、支持硬件、变体、量化、策略、精度标签、HF 模型仓库映射、占位符、Docker 镜像、交互式面板功能定义等，共 1260 行。
2. 新增基准数据文件 `inkling-small-benchmarks.jsx`: 导出 `benchmarks` 数组，为各硬件 / 变体 / 量化 / 策略 / 节点组合记录已验证的 SGLang 版本与精度指标。
3. 新增 cookbook 页面 `Inkling-Small.mdx`: 引入上述配置与基准组件，提供安装说明、交互式部署面板、curl 示例和额外部署提示。
4. 修改部署组件 `_deployment.jsx`: 将投机解码提示从仅判断 `--speculative-algorithm` 是否存在并硬编码为 "MTP"，改为动态检测实际算法值并映射为友好名称（如 EAGLE->MTP, DSPARK->DSpark），同时新增 `pinned` 值提示（`specPinnedHint`）。
5. 更新 Inkling 配置 `inkling.jsx`: 将 Docker 镜像从 `inkling-cu13` 等统一为 `dev-inkling-dspark` 系列，简化 DSpark 备注，DSpark 配方中添加 `--enable-multimodal` 标志。
6. 更新导航配置 `docs.json`: 在 Thinking Machines 分组下添加 Inkling-Small 页面条目。
7. 微调 `Inkling.mdx` 页面（无实质内容变化）。

`docs_new/src/snippets/configs/thinkingmachines/inkling-small.jsx`

新增 Inkling-Small 模型的核心配置数据，包含所有硬件、变体、量化、策略、精度标签、HF 模型仓库映射、Docker 镜像及交互式面板定义，共计 1260 行。

```
// Inkling-Small 资源配置
// 此文件导出供 Deployment 组件使用的配置对象
export const config = {
  modelName: "Inkling-Small",
```

```
// 支持的硬件平台（结构与 Inkling 相同系列）
supportedHardware: [
  "h200", "b200", "b300", "gb200", "gb300",
  "mi350x", "mi355x",
],

// 变体：默认与 LoRA
variants: [
  { id: "default", label: "Default" },
  { id: "lora", label: "LoRA" },
],

// 量化：NVFP4 与 BF16
quantizations: [
  { id: "nvfp4", label: "NVFP4" },
  { id: "bf16", label: "BF16" },
],

// 策略：每种硬件提供一个 verified 的 balanced 配方
strategies: [
  { id: "balanced", label: "Balanced" },
  { id: "mtp", label: "MTP" },
  { id: "dspark", label: "DSpark" },
  { id: "long_context", label: "Long Context (MXFP8 KV)" },
],

// 节点选项
nodesOptions: [
  { id: "single", label: "Single Node" },
  { id: "multi-2", label: "Multi-Nodes" },
],

// 精度标签数组（用于表格展示）
accuracyLabels: [
  ["gsm8k_pct", "GSM8K", "%"],
  ["bfcl_pct", "BFCL (EXACT)", "%"],
  ["mmau_pct", "MMAU", "%"],
  ["mmmu_pro_pct", "MMMU-Pro", "%"],
  ["aime25_pct", "AIME25 (pass@1)", "%"],
  ["aime26_pct", "AIME26 (pass@1)", "%"],
  ["niah_512k_pct", "NIAH @512K", "%"],
  ["niah_1m_pct", "NIAH @1M", "%"],
  ["hle_pct", "HLE", "%"],
],

// HF 模型仓库映射：variantlquant -> repo name
modelName: {
  "defaultlnvfp4": "thinkingmachines/Inkling-Small-NVFP4",
  "defaultlbf16": "thinkingmachines/Inkling-Small",
}
```

```

    "loralnvfp4": "thinkingmachines/Inkling-Small-NVFP4",
    "loralbf16": "thinkingmachines/Inkling-Small",
  },
  // ... 后续 placeholder、curl、dockerImages、playgroundFeatures 等
};

```

docs_new/src/snippets/_deployment.jsx

修改部署组件，将投机解码提示从硬编码 'MTP' 改为动态检测算法名称，提升通用性，支持 MTP、DSpark 等。

```

// 检测投机解码算法名称，用于提示横幅
const specAlgoFlag = effFlags.find(
  (f) => f.split(/\s=/)[0] === "--speculative-algorithm");
const specMrrFlag = effFlags.find(
  (f) => f.split(/\s=/)[0] === "--max-running-requests");

// 两种情况：标志缺失（SGLang 静默限制 48）或已固定
const mtpHint = !!specAlgoFlag && !specMrrFlag;
const specPinnedHint = !!specAlgoFlag && !!specMrrFlag;
const specMrrValue = specMrrFlag
  ? (specMrrFlag.split(/\s=/).filter(Boolean)[1] || "")
  : "";

// 算法标签映射：将 --speculative-algorithm 的值转换为用户友好名称
const SPEC_ALGO_LABEL = {
  EAGLE: "MTP", EAGLE3: "MTP", FROZEN_KV_MTP: "MTP",
  DSPARK: "DSpark", DFLASH: "DFlash", NGRAM: "N-gram",
  STANDALONE: "standalone draft",
};

// 生成算法名称，默认兜底为 "MTP"
const specAlgoName = () => {
  if (!specAlgoFlag) return "MTP";
  const v = specAlgoFlag.split(/\s=/).filter(Boolean)[1] || "";
  return SPEC_ALGO_LABEL[v.toUpperCase()] || v || "MTP";
}();

```

docs_new/src/snippets/configs/thinkingmachines/inkling-small-benchmarks.jsx

新增 Inkling-Small 基准精度数据，覆盖各硬件和策略组合，用于部署面板的准确度展示。

```

// 基准数据数组：每个条目按 (hw, variant, quant, strategy, nodes) 匹配配置，
// 提供 accuracy 指标，key 对应 config.accuracyLabels
export const benchmarks = [
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
    sglang_version: "dev-inkling-dspark (b7252cc)",
    accuracy: { aime26_pct: 95.42, bfcl_pct: 76.54, mmau_pct: 76.30 } },
  { match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
    sglang_version: "dev (cb12a15)",

```

```
    accuracy: { gsm8k_pct: 96.29 } },  
    // ... 更多条目  
];
```

评论区精华

无显著讨论。PR 由 ispobock 提交，zijiexia 审核批准，无 review 评论。

风险与影响

- 风险：_deployment.jsx 的改动影响所有使用该组件的 cookbook 页面，需验证新逻辑覆盖所有投机解码算法；Docker 镜像标签统一可能让现有用户感到困惑；新增配置数据需保持键对齐。
- 影响：Inkling-Small 用户获得官方部署指南；团队需维护此文档，但组件通用化已减轻未来添加新模型的工作量。

关联脉络

本次 PR 与现有 Inkling cookbook 属于同一模型系列，后续可能会继续添加其他规格的模型。文档组件 (_deployment.jsx) 的改进也为未来添加新模型提供了复用基础。