

PR #36286 完整报告

sgl-project/sglang

docs(cookbook): add IBM Granite 4.2 cookbook

合并时间: 2026-08-25 22:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36286>

执行摘要

该 PR 为 IBM Granite 4.2 系列新增完整 cookbook 文档，包括部署配置、性能基准、页面入口和供应商 logo，全部为文档与静态资源变更，不涉及运行时逻辑，风险低。

功能与动机

PR body 明确目的是在 cookbook 自回归网格中展示 IBM 品牌，并提供 Granite 4.2 的验证部署路径。作者通过 `mint validate` 与 `mint broken-links` 校验链接完整性，确保文档质量。

实现拆解

1. 新增 docs/cards/logos/ibm.png: 940x525 透明背景 PNG，包含 IBM 8-bar logo，强制添加以绕过 *.png 的 gitignore 规则。
2. 新增 docs/src/snippets/configs/ibm-granite/granite-4.2.jsx: 导出 config 对象，定义 3B/8B/30B 变体、BF16 量化、模型名映射、启动命令模板、Playground 特性（TP、推理解析器、PD 分离、HiCache）以及验证过的 cells。
3. 新增 docs/src/snippets/configs/ibm-granite/granite-4.2-benchmarks.jsx: 导出 benchmarks 数组，覆盖 H200/B200 上各变体的 TTFT、TPOT 和吞吐数据，供 Deployment 组件渲染。
4. 新增 docs/cookbook/autoregressive/IBM/Granite-4.2.mdx: 页面内容包含安装指南、部署命令生成器、模型介绍（上下文长度 131,072，Apache-2.0 协议）和基准说明。
5. 修改 docs/docs.json: 在自回归 cookbook 导航中插入 IBM 分组，指向新页面。
6. 修改 docs/cookbook/autoregressive/intro.mdx: 在厂商卡片网格中添加 IBM 卡片，链接到 Granite-4.2 页面。

所有变更均为新增或简单的配置挂接，无运行时代码改动。

docs/src/snippets/configs/ibm-granite/granite-4.2.jsx

该文件定义 Deployment 组件所需的完整配置，决定启动命令与 Playground 选项。

```
// 该文件导出 Deployment 组件所需的模型配置。
// variants 定义模型变体，modelName 映射到实际 Hugging Face 仓库。
export const config = {
  modelName: "Granite 4.2",
  supportedHardware: ["h200", "b200"],
  variants: [
    { id: "3b", label: "3B", subtitle: "Dense" },
```

```

    { id: "8b", label: "8B", subtitle: "Dense" },
    { id: "30b", label: "30B", subtitle: "Dense" },
  ],
  quantizations: [{ id: "bf16", label: "BF16" }],
  modelNames: {
    "3blbf16": "ibm-granite/granite-4.2-3b",
    "8blbf16": "ibm-granite/granite-4.2-8b",
    "30blbf16": "ibm-granite/granite-4.2-30b",
  },
  // cells 定义了特定硬件和模型组合下验证过的启动命令。
  cells: [
    {
      match: { hw: "h200", variant: "3b", quant: "bf16", strategy: "balanced", nodes: "single" },
      verified: true,
      flags: [
        "--model-path {{MODEL_NAME}}",
        "--tp 1",
        "--mem-fraction-static 0.8",
        // Granite 4.2 通过 nemotron_3 推理解析器支持思考模式
        "--reasoning-parser nemotron_3",
        "--tool-call-parser qwen3_coder",
        "--host {{HOST_IP}}",
        "--port {{PORT}}",
      ],
    },
  ],
  // 其他变体类似，此处省略。
];

```

docs/src/snippets/configs/ibm-granite/granite-4.2-benchmarks.jsx

提供 Granite 4.2 在 H200/B200 上的性能数据，支撑 Deployment 组件的基准展示。

```

// 基准数据按硬件与模型变体组织。
// 每项包含在指定 workload 下的首 token 时延、每 token 时延和吞吐。
export const benchmarks = [
  {
    match: { hw: "h200", variant: "3b", quant: "bf16", strategy: "balanced", nodes: "single" },
    sglang_version: "d59c1ddf7",
    speed: [
      {
        workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 131.17,
        tpot_ms: 3.51,
        tokens_per_sec_per_gpu: 2472.96,
      },
      {
        workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 1092.54,
        tpot_ms: 7.23,
      },
    ],
  },
];

```

```
tokens_per_sec_per_gpu: 17416.41,  
    },  
  ],  
},  
// 其他硬件与变体组合类似，此处省略。  
];
```

评论区精华

该 PR 只有一条 review 评论（zijiexia 的 Approve），无实质性讨论。作者通过 issue 评论 [/tag-and-rerun-ci](#) 重跑 CI，表明关注 CI 通过。最终 PR 由作者自行合并。

风险与影响

主要风险：

- 部署命令和基准数据固定于特定 SGLang commit（d59c1ddf7、d10a656ad8），后续版本演进可能使命令或数据过期。
- logo 为强制添加，未来若清理 gitignore 规则可能影响文件跟踪。
- 内部链接依赖目录结构，文档重构时需同步更新。

影响：

- 用户：获得 Granite 4.2 的权威部署指南和性能参考。
- 系统：无运行时影响。
- 团队：新增一个完整的 cookbook 模板，便于后续模型文档复用。

关联脉络

该 PR 与近期 Kimi-K2.7-Code cookbook（#36246）和 MiniMax-H3 文档更新（#36169）同属 [docs/](#) 下的持续演进，展示了新增模型 cookbook 的标准流程：添加 logo → 编写配置脚本 → 补充基准 → 挂接导航与入口。这可能是一个更大的“模型 cookbook 体系”建设的一部分。