PR #36827 完整报告

sgl-project/sglang

[Docs] Add GLM-5.3 cookbook

合并时间：2026-08-28 22:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36827

执行摘要

该 PR 为 Z.ai 的 GLM-5.3 新增官方部署 cookbook，复用 GLM-5.2 的部署骨架，覆盖 H200、B200、GB300、B300 与 AMD MI300X/MI325X/MI355X 上 FP8/BF16（及 Experimental NVFP4）的部署矩阵，并附带 15 个经实测标记 verified 的 NVIDIA 单节点 benchmark 单元格。页面已接入 docs.json 导航并带 NEW 标签。需要特别说明：PR body 描述的是“纯文档占位、移除 NVFP4 与实测数据”的早期状态，而 12 个 commit 演进后的合并结果恰恰包含了这些内容，文档消费方应以页面实际渲染为准。

功能与动机

PR body 的动机很明确：GLM-5.3 keeps the GLM-5.2 base architecture, so the existing deployment recipes can be reused with the released GLM-5.3 checkpoints。也就是 GLM-5.3 与 GLM-5.2 同架构，部署配方可以直接复用，只需对齐新发布的 checkpoint：FP8 用 zai-org/GLM-5.3、BF16 用 zai-org/GLM-5.3-BF16。同时要同步 release 仓库的新默认值（temperature=1.0、top_p=0.95）、reasoning-effort 指引、clear_thinking 行为，并澄清 DSA 部署约束（LayerSplit 依赖带 prefill context parallelism 的 Mooncake PD prefill worker）。后续 commit 又补充了实测性能数据与 NVFP4 实验性单元格，使文档从“纯占位”升级为“可用配方 + 实测参考”。

实现拆解

1. 新增 cookbook 页面docs/cookbook/autoregressive/GLM/GLM-5.3.mdx：front matter 声明 title: GLM-5.3、description 与 tag: NEW；正文依次渲染安装 Accordion、部署矩阵、Playground，并导入 /src/snippets/_deployment.jsx、glm-5.3.jsx、glm-5.3-benchmarks.jsx 完成交互式命令生成。页面用 Warning 注明 DSA indexer top-k 目前只在 --dsa-topk-backend sgl-kernel 上验证。
2. 新增部署配置数据源docs/src/snippets/configs/zai-org/glm-5.3.jsx（1028 行）：导出单一字面量 config，定义硬件列表、量化选项、checkpoint 映射、命令占位符、benchmarkCommands、各硬件 dockerImages 以及 Playground 的 Attention CP knob（DSA CP 在 ROCm 与多节点场景被 disable）。关键设计是单元格反规范化：不写 --nnodes、--node-rank、--host、--port 字面量，由引擎在渲染时注入，避免多节点复制出错。
3. 新增 benchmark 数据源glm-5.3-benchmarks.jsx（221 行）：用 match 元组（hw/variant/quant/strategy/nodes）与部署单元格一一对应；15 个 NVIDIA 单节点单元格被实测并标记 verified，notes 字段说明速度数据在 SGLANG_SIMULATE_ACC_LEN 固定

接受长度下测得，仅代表吞吐机制而非正确性证据。NVFP4 单元格来自第三方 RadixArk/GLM-5.3-NVFP4，部分标记为 *experimental / pending*。

4. 导航接入与配置校验docs/docs.json（+1 行）：在 GLM 分组把新页面插到 GLM-5.3-Flash 之前，并依赖 mint validate、broken-links 和部署矩阵一致性检查。
5. 数据决策演进：FP8 balanced 配方经吞吐对比后提升为 high-throughput（如 H200 上 2558 vs 1762 tok/s/GPU）；NVFP4 先加入验证单元格后整体退回 Experimental；最后一个 commit 把准确率指标切换为 AIME26 并补录 NVFP4 分数。

docs/src/snippets/configs/zai-org/glm-5.3.jsx

GLM-5.3 部署矩阵的唯一数据源，1028 行新增配置定义硬件、量化、checkpoint 映射、复现命令和 Playground knob，是整个 cookbook 的核心契约。

```
// GLM-5.3 部署配置的核心契约（docs/src/snippets/configs/zai-org/glm-5.3.jsx）
// Mintlify 在 hydration 时会对该字面量重新求值，因此不能用 spread / 调用 / IIFE；
// 单元格是反规范化的，--nnodes、--node-rank、--host、--port 等由引擎注入。
export const config = {
  modelName: 'GLM-5.3',

  // 硬件矩阵覆盖 NVIDIA H200/B200/GB300/B300 与 AMD MI300X/MI325X/MI355X。
  supportedHardware: ['h200', 'b200', 'gb300', 'b300', 'mi355x', 'mi325x', 'mi300x'],

  // 官方只发布单一 checkpoint，不做 size/mode 拆分；
  // FP8 与 BF16 走官方仓库，NVFP4 来自第三方 RadixArk，标记为 Experimental。
  variants: [{ id: 'default', label: 'GLM-5.3', subtitle: 'MoE • DSA' }],
  quantizations: [
    { id: 'fp8', label: 'FP8' },
    { id: 'bf16', label: 'BF16' },
    { id: 'nvfp4', label: 'NVFP4 (Experimental)' },
  ],
  modelNames: {
    'defaultIfp8': 'zai-org/GLM-5.3',
    'defaultIbf16': 'zai-org/GLM-5.3-BF16',
    'defaultInvfp4': 'RadixArk/GLM-5.3-NVFP4',
  },

  // 命令占位符：页面按用户选择的单元格替换后生成可执行命令。
  placeholders: {
    HOST_IP: { target: 'command', label: 'Bind host', default: '0.0.0.0' },
    PORT: { target: 'command', label: 'Bind port', default: '30000' },
    NODE0_IP: { target: 'command', label: 'Head node IP', default: '<node0-ip>' },
    NODE_RANK: { target: 'command', label: 'This node rank', default: '<node-rank>' },
    CURL_HOST: { target: 'curl', label: 'Server host', default: 'localhost' },
    CURL_PORT: { target: 'curl', label: 'Server port', default: '30000' },
  },

  // Playground 中 Attention Parallelism 卡片的 DSA CP knob：
  // DSA prefill CP（context parallelism）在 H200/B200/GB300/B300 上可用，
  // 但 ROCm 上该路径尚未验证，因此对 MI300X/MI325X/MI355X 直接禁用。
```

```
  playgroundFeatures: {
    attention: {
      knobs: [
        { id: 'tp', label: 'TP', values: [null, 4, 8] },
        { id: 'cp', label: 'CP (DSA prefill)', values: [null, { value: 1, label: 'Off' }, 4, 8],
          disable: [
            { when: { hw: ['mi355x', 'mi325x', 'mi300x'] },
              reason: 'ROCm DSA-CP 路径尚未在 AMD 上验证，保持 CP 关闭。' },
            // 另有针对 multi-2 节点的 disable 条件（reason 原文不可见，省略）。
          ] },
      ],
    },
  },
};
```

docs/src/snippets/configs/zai-org/glm-5.3-benchmarks.jsx

存放 GLM-5.3 的全部 benchmark 条目，用 match 元组绑定部署单元格，并承载 verified/experimental/pending 三级数据标注；包含 SGLANG_SIMULATE_ACC_LEN 机制验证说明。

```
// GLM-5.3 benchmark 数据源（docs/src/snippets/configs/zai-org/glm-5.3-benchmarks.jsx）
// 每个条目用与 glm-5.3.jsx 部署单元格一致的 match 元组作键；
// 只有 match 而没有数据体的条目渲染为 pending，补上实测结果后才对外展示。
export const benchmarks = [
  {
    // H200 + FP8 + Low-Latency + 单节点：对应聊天场景的默认单元格。
    match: { hw: 'h200', variant: 'default', quant: 'fp8', strategy: 'low-latency', nodes: 'single' },
    sglang_version: 'main @ 20a491d1d311', // 数据与具体 commit 绑定，版本更新后需复测。
    accuracy: { gsm8k_pct: 97.42 },
    speed: [
      { workload: { dataset: 'random', isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 780, tpot_ms: 3.71, tokens_per_sec_per_gpu: 252 },
      { workload: { dataset: 'random', isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 5499, tpot_ms: 14.20, tokens_per_sec_per_gpu: 920 },
    ],
    notes:
      '速度数据在 SGLANG_SIMULATE_ACC_LEN=3.5 下用 EAGLE 5/1/6 draft 测得；' +
      '固定接受长度只验证吞吐机制，不能作为正确性证据，准确率数据不携带该环境变量。',
  },
  // 其余已验证单元格与 NVFP4 experimental / pending 项结构相同，依此类推。
];
```

评论区精华

- mintlify[bot]：> Preview deployment for your docs … ✅ Ready … View Preview。仅提供预览链接，无技术讨论。
- zijiexia 的 commit 说明：> A throughput sweep on sglang main @ 20a491d1d311 … shows the balanced recipe beating the high-throughput one on total tok/s/GPU on

every FP8 box: 2558 vs 1762 on H200, 5025 vs 3997 on B200, 5259 vs 4067 on B300。这是将 FP8 balanced 升格为 high-throughput 的依据。

- JustinTong0323 在 commit 中先把 NVFP4 单元格加入并验证，随后退回 experimental，理由是官方未发布 FP4 权重。
- 无人工 review 评论，最终由 JustinTong0323 APPROVED。

风险与影响

- 数据时效性：benchmark 全部绑定 sglang main @ 20a491d1d311，引擎变化后数字会失效，文档缺少过期提示。
- 误用风险：SGLANG_SIMULATE_ACC_LEN 固定接受长度产生的速度数字被标注为机制验证，用户可能误当真实性能。
- 第三方权重：NVFP4 指向 RadixArk/GLM-5.3-NVFP4，上游变更会导致文档断链。
- 文档一致性：PR body 与实际内容矛盾，会让后续维护者与 changelog 工具困惑。
- 影响面：仅 docs 域，对用户选型与部署有直接帮助，对团队意味着后续数据维护负担。

关联脉络

该 PR 与近期 cookbook 系列在模式上高度一致：36804 新增 Hy4-Preview 页面、36808 做 follow-up 校准配方、36823 做页面重命名，说明 SGLang 文档团队已形成“新增页面 → 实测补数 → 持续校准”的成熟流程。GLM-5.3-Flash 与 GLM-5.2 已存在于 docs.json 导航中，本次新增的 GLM-5.3 是 GLM 家族文档的延续；而 benchmark 数据标注方式（match 键、verified/experimental/pending、环境变量说明）也与 Hy4-Preview 的配置 snippet 风格一致，未来大概率会有 follow-up PR 更新实测数据或修正 NVFP4 状态。