PR #36440 完整报告

sgl-project/sglang

Add GLM-5.3-Flash cookbook

合并时间：2026-08-26 22:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36440

执行摘要

本 PR 为 GLM-5.3-Flash 添加了完整的部署 cookbook，覆盖 GB300 / H100 / H200 / B200 / B300 / GB200 六种硬件，内置 Low Latency（adaptive MTP 5/1/6）与 High Throughput（关闭推测解码）两种策略，并支持 KV/DSA 后端、VLM 传输与 HiCache 的叠加配置。同时新增 GB300 实测基准数据，并对部署面板 _deployment.jsx 做了通用性增强：URL hash 恢复时，被隐藏或禁用的 overlay 选项会自动回退到第一个可用项。整体是文档与前端 snippet 变更，不涉及 SGLang 运行时。

功能与动机

SGLang 的 cookbook 体系持续为热门模型提供 Day-0 部署指南，让用户不必自行摸索命令行参数。PR body 明确指出目标是“deployment recipes for GB300 / H100 / H200 / B200 / B300 / GB200 / MI300X with NEXTN speculative decoding and multimodal serving”。GLM-5.3-Flash 作为新一代混合注意力 MoE 模型（MLA + DSA + KDA、288 路由专家、原生 MTP 草稿层），需要一份可复现的部署文档来承载这些新特性。

实现拆解

1. 配置声明：docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx 中的 config 对象声明了 supportedHardware、matchDims（strategy）与 overlayDims（kvDsaPair / mmTransport / hicache），每个 overlay 选项携带 flags、stripPrefixes、disableReason 等元数据，部署面板据此生成命令并控制选项可用性。
2. 基准数据：glm-5.3-flash-benchmarks.jsx 提供 benchmarks 数组，GB300 上 Low Latency 与 High Throughput 各有真实测量值，High Throughput 覆盖 16 / 64 / 256 三档并发；其余硬件留空占位。注释中特别区分了“模拟 accept length 的吞吐数据”与“非模拟的 GSM8K 准确率”，避免误导用户。
3. 文档页面：GLM-5.3-Flash.mdx 导入 Deployment 与 Playground，先用自然语言说明两种策略差异与 Verified 徽标语义，再通过组件呈现交互式配置。模型介绍部分给出了 320B 总参 / 18B 激活、45 层文本层与 24 层视觉编码器等关键架构信息。
4. 面板通用逻辑增强：_deployment.jsx 修改 validateSelection，使 overlay 选项在 showWhen 隐藏或规则禁用时，从“保留非法值”变为“回退到第一个可用选项”，与用户手动重新选择的行为一致；selSummary 改为过滤空字段后拼接，避免缺少维度时出现空分隔符。
5. 路由与清理：docs/docs.json 在 GLM 分组注册新页面；GLM-5.2.mdx 删除 tag: NEW，防止新旧模型同时显示“新”标记。

6. 测试与部署配套：无新增自动化测试，属于纯文档维护；CI 中 PR Test 通过，Extra Run 失败未在讨论中展开。

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

提供 GB300 实测吞吐与 GSM8K 准确率数据，其余硬件留占位，是部署面板基准矩阵的数据源。

```
// 基准矩阵数据：Deployment 面板按 hw + strategy 匹配单元格。
// 只有真实测过的条目才填 speed/accuracy，其余为占位符，
// 页面会据此显示 Verified / Not Verified 徽标。
export const benchmarks = [
  {
    match: { hw: "gb300", strategy: "low-latency" },
    sglang_version: "f13cb6f6a7",
    latencyPercentile: "Mean",
    speed: [
      {
        workload: { dataset: "random", isl: 1024, osl: 256, max_concurrency: 16, num_prompts:
        80 },
        ttft_ms: 589.2,
        tpot_ms: 6.48,
        tokens_per_sec_per_gpu: 2277.46,
      },
    ],
    accuracy: { gsm8k_pct: 97.50 },
    // 注释明确说明 SGLANG_SIMULATE_ACC_LEN=3 是吞吐机制验证，
    // 不是准确率数据来源；准确率来自共享的非模拟 GSM8K gate。
    notes:
      "Measured on 4x GB300 (TP4/EP4) with the final weights (zai-org/GLM-5.3-Flash,
      c5b82b63e37b) at the rc2 cut (f13cb6f6a7), adaptive MTP 5/1/6 with SGLANG_SIMULATE_
      ACC_LEN=3 ...",
  },
  {
    match: { hw: "gb300", strategy: "high-throughput" },
    sglang_version: "f13cb6f6a7",
    latencyPercentile: "Mean",
    speed: [
      { workload: { dataset: "random", isl: 1024, osl: 256, max_concurrency: 16, num_prompts:
      80 }, ttft_ms: 684.63, tpot_ms: 11.53, tokens_per_sec_per_gpu: 1410.4 },
      { workload: { dataset: "random", isl: 1024, osl: 256, max_concurrency: 64, num_prompts:
      320 }, ttft_ms: 1691.64, tpot_ms: 19.88, tokens_per_sec_per_gpu: 3023.31 },
      { workload: { dataset: "random", isl: 1024, osl: 256, max_concurrency: 256, num_prompts:
      1280 }, ttft_ms: 5192.23, tpot_ms: 43.35, tokens_per_sec_per_gpu: 4856.19 },
    ],
    accuracy: { gsm8k_pct: 97.50 },
    notes:
      "Measured on 4x GB300 (TP4/EP4) with the final weights ..., speculative decoding off,
      after two discarded warmups per row ...",
  },
```

```
  // 其他硬件暂无实测数据，保留空单元格以便后续 PR 补充。
  { match: { hw: "h100", strategy: "low-latency" } },
  { match: { hw: "h100", strategy: "high-throughput" } },
  { match: { hw: "h200", strategy: "low-latency" } },
  { match: { hw: "h200", strategy: "high-throughput" } },
  { match: { hw: "b200", strategy: "low-latency" } },
  { match: { hw: "b200", strategy: "high-throughput" } },
  { match: { hw: "b300", strategy: "low-latency" } },
  { match: { hw: "b300", strategy: "high-throughput" } },
  { match: { hw: "gb200", strategy: "low-latency" } },
  { match: { hw: "gb200", strategy: "high-throughput" } },
];
```

docs/src/snippets/_deployment.jsx

部署面板通用组件：修改 validateSelection 使 overlay 维度在 hash 选中的选项被隐藏或禁用时正确回落，修改 selSummary 拼接逻辑，影响所有 cookbook 的 URL 恢复行为。

```
// 将 URL hash 中解析出的选择（可能已过期）“吸附”到真实可用的单元格。
// 主维度逐级校验；overlay 维度不参与单元格匹配，但可能被 showWhen 隐藏
// 或被规则禁用，因此需要像用户手动重新选择一样做一次可用性过滤。
const validateSelection = (cells, parsed) => {
  const valid = {};
  // 先处理主维度：保持优先级顺序，逐维寻找与已确定维度一致的单元格。
  for (const dim of DIMENSIONS) {
    const want = parsed[dim];
    const works = cells.some(
      (c) =>
        c.match[dim] === want &&
        DIMENSIONS.slice(0, DIMENSIONS.indexOf(dim)).every((d) => c.match[d] === valid[d]),
    );
    if (works) {
      valid[dim] = want;
    } else {
      const fallback = cells.find((c) =>
        DIMENSIONS.slice(0, DIMENSIONS.indexOf(dim)).every((d) => c.match[d] === valid[d]),
      );
      valid[dim] = fallback ? fallback.match[dim] : want;
    }
  }
  // Overlay 维度跟随主选择，但 hash 可能给出一个被隐藏或禁用的选项；
  // 此时应该回退到当前组合下第一个可用选项，而不是保留非法值。
  for (const spec of overlayDimSpecs) {
    const want = parsed[spec.id];
    const opts = spec.options || [];
    const picked = opts.some((o) => o.id === want)
      ? want
      : spec.default ?? (opts[0] && opts[0].id) ?? "";
    const withPick = { ...valid, [spec.id]: picked };
    const usable = visibleOptions(spec, withPick).filter((o) => !optionDisabled(o, withPick));
```

```
      // 如果 picked 不可见或不可用，则取第一个可用的选项；
      // 若完全没有可用项，则保留 picked（防御性兜底）。
      valid[spec.id] = usable.some((o) => o.id === picked)
        ? picked
        : (usable[0] && usable[0].id) ?? picked;
    }
    return valid;
  };

  // 基准命令弹窗中的摘要拼接：只输出非空字段，
  // 避免缺少 variant 或 quant 维度时出现“··”之类的空分隔符。
  const selSummary = [
    sel.hw && sel.hw.toUpperCase(),
    sel.variant,
    sel.quant && sel.quant.toUpperCase(),
    sel.strategy,
    sel.nodes,
  ]
    .filter((part) => part !== undefined && part !== null && part !== "")
    .join(" · ");
```

评论区精华

本 PR 没有实质性的 review 讨论。唯一审核人 zijiexia 直接 APPROVED，Mintlify bot 提供了预览部署链接。因此没有需要提炼的技术交锋。

风险与影响

- 部署面板回归风险：_deployment.jsx 是多个 cookbook 共用的组件，validateSelection 的行为变化可能影响既有页面在特定 URL hash 下的默认回退结果，需要关注面板的视觉回归。
- 基准数据时效性：GB300 数据固定到特定 commit 与权重快照，后续 SGLang 版本演进后可能过期；其他硬件仍为空占位，用户可能误以为未验证。
- 配置约束依赖人工提示：FP8 在 Hopper 禁用、HiCache L3 需要 Mooncake 等约束只在 UI 文案中体现，没有自动化校验，配置错误只能靠用户阅读。

关联脉络

本 PR 属于 sglang 文档 cookbook 系列的扩展：与 #36496（Qwen3.8-Flash-Next cookbook）共用同一套配置与部署面板模式；#36499 展示了 cookbook 上线后配套模型支持 PR 的文档修正方式。GLM-5.3-Flash cookbook 中涉及的 DSA 稀疏注意力、HiCache、FP8 KV cache 等选项，也与 sglang 在 kv-cache、hicache、quant 方向的演进一致，后续可关注模型正式支持的 PR 与更完整硬件基准的补充。