# PR #28448 完整报告

sgl-project/sglang

docs(cookbook): tune GLM-5.2 MTP to 5-1-6 and simplify launch flags

合并时间：2026-06-17 01:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28448

## 执行摘要

本 PR 调整了 GLM-5.2 cookbook 的 MTP 配置（从 3-1-4 改为 5-1-6）并简化了启动参数，同时更新了多款硬件的性能基准。通过利用模型 MTP 头的高接受率，低延迟场景吞吐提升 31-34%、TPOT 降低 18-28%。

## 功能与动机

- GLM-5.2 的 MTP 头很强，低并发下接受长度接近饱和（GB300 上约 6/6），因此更长的 draft（5 个步骤、6 个草稿 token）能显著提升性能。
- 模型在 transformers 5.8.1+ 中原生集成，且权重公开，无需 --trust-remote-code 和 --env HF_TOKEN，移除这些标志可降低用户部署时的认知负荷。

## 实现拆解

1. MTP 参数调整：在 glm-5.2.jsx 中，将 low-latency 的 speculative.options 从 mtp-314 改为 mtp-516；balanced 选项保持 1-1-2 并添加 (balanced) 标签。同步更新所有 low-latency cell 的 flags。
2. 标志简化：从所有 cell flags 中移除 --trust-remote-code；从 placeholders 中移除 HF_TOKEN；并修改共享渲染组件 _deployment.jsx，使其只在配置声明 HF_TOKEN 占位符时才注入该环境变量。
3. 基准更新：在 glm-5.2-benchmarks.jsx 中更新 H200/B200/GB300 的 low-latency 性能数据，并添加注释说明提升百分比。
4. 文档补充：在 GLM-5.2.mdx 中新增 DSA KV-cache 自动默认段落，解释 Blackwell/Hopper 上 KVCache dtype 的选择逻辑。

### docs_new/src/snippets/configs/zai-org/glm-5.2.jsx

核心配置，调整 MTP 方案并简化标志

```
// …
speculative: {
  options: [
    { id: "current", label: "Inherited from base" },
    { id: "off", label: "Off (greedy)" },
    // 低延迟方案：5 个推理步骤，6 个草稿 token
    { id: "mtp-516", label: "EAGLE / MTP 5-1-6 (low-latency)",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 5",
```

```
            "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 6"] },
    // 均衡方案：1 个推理步骤，2 个草稿 token
    { id: "mtp-112", label: "EAGLE / MTP 1-1-2 (balanced)",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 1",
            "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 2"] },
  ],
},
// ...
// H200 low-latency cell 示例 flags（去除了 --trust-remote-code 和 HF_TOKEN）
{
  match: { hw: "h200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" },
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 5",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 6",
    "--mem-fraction-static 0.8",
    "--cuda-graph-max-bs 32",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

## docs_new/src/snippets/_deployment.jsx

共享部署渲染组件，条件化 HF_TOKEN env 变量

```
// 从 config.placeholders 中检查是否声明了 HF_TOKEN 占位符
// 仅当存在时才注入 --env "HF_TOKEN=..."，从而允许公开模型（如 GLM-5.2）免填 token
const dockerLines = [
  "docker run --gpus all",
  "  --shm-size 32g",
  multinode ? "  --network host" : `  -p ${servePort}:${servePort}`,
  "  -v ~/.cache/huggingface:/root/.cache/huggingface",
  // 关键变更：原本硬编码为 "--env \"HF_TOKEN={{HF_TOKEN}}\"" ,
  // 现在改为有条件添加
  ...(config.placeholders && config.placeholders.HF_TOKEN
    ? [`  --env "HF_TOKEN={{HF_TOKEN}}"`]
    : []),
  ...cellEnv.map((e) => `  --env ${e}`),
  "  --ipc=host",
  `  ${image}`,
  "  sglang serve",
  ...flags.map((f) => "    " + f),
];
```

## 评论区精华

无实质性讨论，PR 被 Reviewer 快速批准。

## 风险与影响

- 风险：共享组件 _deployment.jsx 的条件逻辑变更可能影响其他模型配置——如果 gated 模型未在 placeholders 中声明 HF_TOKEN，则部署命令会缺少 token 环境变量。本次修改已确保所有使用该组件的配置均正确声明（PR body 说明 gated models unaffected），但后续新增模型配置时需注意这一约束。
- 影响：正面影响明确——用户部署 GLM-5.2 可获得更优的默认低延迟性能（吞吐 +31~34%，TPOT -18~28%），且命令行更简洁；开发者可复用条件渲染模式以区分公开与门控模型。

## 关联脉络

本 PR（#28448）是 #28437（GLM-5.2 deployment cookbook）的后续优化，基于实际测量数据调整 MTP 参数并简化部署流程。后续可能继续添加 B300 等更多硬件基准。