PR #33935 完整报告

sgl-project/sglang

Clean GLM-5.2 NVFP4 cookbook

合并时间：2026-08-07 11:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33935

执行摘要

本 PR 是纯文档清理：从 GLM-5.2 NVFP4 低延迟配置示例中移除 B300/GB300 上两个冗余启动参数（--kv-cache-dtype fp8_e4m3 与 --bf16-gemm-backend cutedsl），使文档示例与 SGLang 在 Blackwell 上的默认行为一致，并统一三个 Blackwell 列的配置风格。不涉及任何代码逻辑变更。

功能与动机

PR body 说明：这两个 flag 是冗余的。SGLang 在 Blackwell（B200/GB300/B300）上会自动为 DSA KV cache 选择 fp8_e4m3（页面 attention notes 已写明），而 cutedsl 是该硬件上的默认 BF16 GEMM 后端。B200 的 NVFP4 配方从未携带这两个参数，删除后三个 Blackwell 列保持一致。核心意图是避免文档示例与文档正文矛盾，防止用户复制多余的参数。

实现拆解

1. 定位配置条目：在 docs/src/snippets/configs/zai-org/glm-5.2.jsx 中找到 hw: "b300" 与 hw: "gb300"、strategy: "low-latency" 的 NVFP4 配置对象。
2. 删除冗余 flag：从 B300 与 GB300 两个条目的 flags 数组中各删除 "--kv-cache-dtype fp8_e4m3" 与 "--bf16-gemm-backend cutedsl" 两行，共删除 4 行，保持两侧结构对称。
3. 一致性校对：与 B200 high-throughput 等其余条目对比，确认没有其他位置残留这两个参数，三个 Blackwell 列现在统一不再显式携带。
4. 测试与 CI：文档类改动不需要新增测试；PR 的 Base 测试通过，Extra 测试失败与文档变更无关。

docs/src/snippets/configs/zai-org/glm-5.2.jsx

唯一修改的文件。删除 B300/GB300 的 NVFP4 低延迟配置示例中两个冗余 flag，使三个 Blackwell 平台列配置风格一致，并与页面 attention notes 的自动选择描述相符。

```
// GLM-5.2 NVFP4 配置数组中的两条低延迟示例（整理后）
// 本次改动删除了 B300 与 GB300 上显式的 `--kv-cache-dtype fp8_e4m3`
// 和 `--bf16-gemm-backend cutedsl`，原因：SGLang 在 Blackwell 上
// 会自动为 DSA KV cache 选择 fp8_e4m3（页面 attention notes 已声明），
// 且 cutedsl 本来就是默认的 BF16 GEMM 后端，显式指定反而与正文矛盾。

// B300 平台的 NVFP4 低延迟配置（single node）
{
  match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single" }
```

```
    ,
    verified: true,
    env: [],
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 8",
      "--quantization modelopt_fp4",
      "--speculative-algorithm EAGLE",
      "--speculative-num-steps 5",
      "--speculative-eagle-topk 1",
      "--speculative-num-draft-tokens 6",
      "--chunked-prefill-size 8192",
      "--mem-fraction-static 0.85",
      "--max-running-requests 16",
      "--cuda-graph-max-bs 16",
      "--max-prefill-tokens 8192",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },

  // GB300 平台的 NVFP4 低延迟配置（single node），与 B300 同步清理
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single"
    },
    verified: true,
    env: [],
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 4",
      "--quantization modelopt_fp4",
      "--speculative-algorithm EAGLE",
      "--speculative-num-steps 5",
      "--speculative-eagle-topk 1",
      "--speculative-num-draft-tokens 6",
      "--chunked-prefill-size 8192",
      "--mem-fraction-static 0.85",
      "--max-running-requests 16",
      "--cuda-graph-max-bs 16",
      "--max-prefill-tokens 8192",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },
```

评论区精华

本 PR 没有实质性的 review 技术讨论。唯一的外部评论是 mintlify[bot] 自动发布的文档预览部署通知，提供了 Mintlify 预览链接，不涉及代码评审意见。审核人 JustinTong0323 直接批

准（APPROVED），无未解决疑虑。

风险与影响

风险极低。被删除的 flag 在目标硬件上均为默认值，用户直接复制示例不会感知行为差异。唯一潜在风险：文档示例与 Blackwell 默认行为绑定，若未来 SGLang 更改默认 KV cache dtype 或默认 BF16 GEMM 后端，示例可能与实际行为偏离；但页面 attention notes 同一处已声明自动选择，保持一致反而降低了长期维护歧义。影响范围仅限文档阅读者，对系统运行与 CI 无影响。

关联脉络

与 ModelOpt FP4 量化支持（相关 PR 33115，为 --quantization modelopt_fp4 提供在线量化和加载链路）以及 cute DSL 后端的性能工作（相关 PR 33650）同属 Blackwell 量化推理的文档 / 功能演进；本次清理使 GLM-5.2 NVFP4 配置示例与默认行为保持一致，降低了用户复制配置时的困惑。