PR #28731 完整报告

sgl-project/sglang

[cookbook] drop redundant serve flags (GLM-5.2) + fix M3 page-size note

合并时间：2026-06-29 13:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28731

执行摘要

本 PR 根据内部反馈清理了 GLM-5.2 和 MiniMax-M3 cookbook 中的冗余 serve flags，删除手动设置的运行时默认值已正确的参数，提高 --max-running-requests 避免人为限流；同时用真实权重重新测量 B200/GB300 的 benchmark 性能并更新文档。整体降低了 cookbook 配置的维护成本，提高了配置的可读性和准确性。

功能与动机

Lianmin 在 2026 年 6 月 19 日的反馈指出："make the default arguments just work — drop hand-set serve flags where the runtime default is correct or better"。因此本 PR 删除了 GLM-5.2 配置中与运行时默认重复的 --cuda-graph-max-bs-decode 等标志，并调整了 --max-running-requests 以允许更高的并发。同时，上游 PR#28976 已实现 --attention-backend fa4 时自动强制 page_size 为 128，因此 MiniMax-M3 文档中要求显式设置 --page-size 128 的说明已过时，需修正。

实现拆解

1. 清理 GLM-5.2 serve flags：在 glm-5.2.jsx 中，删除所有 cell 中的 --cuda-graph-max-bs-decode 标志，低延迟和高吞吐策略不再显式设置 --max-running-requests，平衡策略提升至 256。为 GB300 高吞吐配置添加 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=512 环境变量。向 GLM-5.2.mdx 添加 benchmark 方法论注释。
2. 修复 MiniMax-M3 page-size 注释：在 minimax-m3.jsx 中删除所有 Blackwell cell 的 --page-size 128，并在 MiniMax-M3.mdx 中修正说明，指出 --attention-backend fa4 已自动强制 page_size 为 128。
3. 更新 benchmark 数据：在 glm-5.2-benchmarks.jsx 中，用 main @ 09ca4fc 上真实权重的测量结果替换旧数据。H200 标记为待重新测量，B200 和 GB300 填入新结果。

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

核心 serve flags 清理：删除冗余的 --cuda-graph-max-bs-decode，调整 --max-running-requests，为 GB300 HT 添加 env SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=512

```
// GLM-5.2 balanced cell (H200 FP8): 去掉 --cuda-graph-max-bs-decode 128,
// 同时将 --max-running-requests 从 80 提升到 256，避免人为限流。
{
```

```
    match: { hw: "h200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" },
    verified: true,
    env: [],
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 8",
      "--dp 8",
      "--enable-dp-attention",
      "--moe-a2a-backend deepep",
      "--speculative-algorithm EAGLE",
      "--speculative-num-steps 1",
      "--speculative-eagle-topk 1",
      "--speculative-num-draft-tokens 2",
      "--mem-fraction-static 0.85",
      // 大 chunked-prefill 是 balanced 策略的主要杠杆；
      // max-running 跟踪 KV 容量（H200 上约 60-80 个 8K+1K 请求）。
      "--chunked-prefill-size 32768",
      "--max-running-requests 256",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  }
```

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

核心 benchmark 数据更新，用真实权重重新测量 B200/GB300，标记 H200 为 pending

```
  // B200 low-latency 策略的新 benchmark 条目
  // 测量条件：8×B200, TP8, 真实权重，--random-range-ratio 1.0, 每次运行 --flush-cache
  // env SGLANG_SIMULATE_ACC_LEN=3.5 固定 EAGLE 接受长度（50% 接受 3/50% 接受 4）
  {
    match: { hw: "b200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" },
    sglang_version: "main @ 09ca4fc",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 757, tpot_ms: 3.22, tokens_per_sec_per_gpu: 32 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 3188, tpot_ms: 9.12, tokens_per_sec_per_gpu: 164 },
    ],
  }
```

评论区精华

本 PR 无实质性 review 讨论。zijiexia 直接批准了变更，机器人评论未提出建设性意见。

风险与影响

- 风险：删除的 flags 可能影响少量已习惯旧配置的用户，但 PR 通过验证确认运行时默认值表现更好。GB300 高吞吐配置的 env 变量是必需的，用户若忽略可能导致 DeepEP 断言失败。benchmark 数据更新可能使用户对比旧版本时产生困惑，但已注明测量条件。

- 影响：cookbook 用户获得更简洁的命令，维护团队减少了冗余配置。无运行时影响。

关联脉络

本 PR 是 cookbook 配置持续优化的一部分。上游 PR#28976（fa4 页面大小自动强制）使 MiniMax-M3 的显式设置不再必要。后续可关注其他模型配置的类似清理。