PR #36660 完整报告

sgl-project/sglang

cookbook: fix GLM-5.3-Flash speculative flag, size Hopper memory, record GSM8K

合并时间：2026-08-27 17:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36660

执行摘要

本 PR 是一次针对 GLM-5.3-Flash cookbook 的文档修正与数据补全：将失效的 --speculative-algorithm NEXTN 统一改为 EAGLE，调整 H100 / H200 的静态显存占比，并为 H100 / H200 / B200 / B300 八个部署单元格补录完整 GSM8K 精度数据、标记为 verified。改动全部位于 docs 目录，不触及运行时代码，影响对象是照着 cookbook 部署 GLM-5.3-Flash 的用户。

功能与动机

commit message 明确指出：low-latency 配方传入的 --speculative-algorithm NEXTN 会被 SpeculativeAlgorithm.from_string 拒绝，枚举成员已折叠进 EAGLE，旧写法会导致用户直接复制文档命令启动失败。这正是本 PR 的第一动因。

PR body 同时提供了完整的 GSM8K 评测结果（全量 1,319 题，zai-org/GLM-5.3-Flash at f040cc72e6，TP8 / EP8），覆盖 8x H100 / H200 / B200 / B300 四类平台、Low Latency 与 High Throughput 两种策略；34 个面板组合全部标记 verified，并明确 HiCache + L3（Mooncake）以及 GB200 / GB300 lane 未测量、保持 unverified。这些数据用于支撑 cookbook 中各推荐配置的 verified 标记。

实现拆解

1. 修正 MTP 启动参数：docs/cookbook/autoregressive/GLM/GLM-5.3-Flash.mdx 第 88 行策略说明与第 168 行示例命令，以及 docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx 中 gb300 / h100 / h200 / b200 / b300 各 low-latency 单元格的 --speculative-algorithm NEXTN 全部改为 EAGLE，消除 from_string 拒绝导致的部署失败。
2. 调整 Hopper 显存占比：h100 low-latency 单元格的 --mem-fraction-static 从 0.75 下调到 0.70，h200 low-latency 单元格显式补上 0.75，以匹配不同 Hopper 显存预算、降低 OOM 风险。
3. 更新验证状态：h100 / h200 / b200 / b300 单元格从 verified: false 或 in-progress 判定改为 verified: true，并用条件式 verificationStatus 精确表达已验证组合——high-throughput 单元格接受 HiCache off 与 L2，low-latency 单元格要求 mmTransport === "auto" 且 hicache === "off"。
4. 补录 GSM8K 基准数据：docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx 将 h100 / h200 / b200 / b300 的八个空 match 条目扩展为带 sglang_version、accuracy.gsm8k_pct 与 notes 的数据条目，记录复现命令 sgl-eval run gsm8k --base-url http://localhost:30000/v1 --num-threads 32 --max-tokens 32768，并注明这

是非 thinking 模式、accuracy-only 数据，与 GB300 行不直接可比。

5. 配套与 CI：无新增测试；文档构建由 Mintlify Preview 验证（bot 评论）。PR Test 通过，PR Test (Extra) 失败，失败原因未在材料中给出，无法判断是否与文档改动相关。

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

核心配置单元：修正所有 low-latency 单元格的 speculative flag（NEXTN -> EAGLE），调整 H100/H200 的 mem-fraction-static，并将验证状态从 in-progress 升级为带条件的 verified。

```
// H100 Low Latency 单元格：调整显存占比、修正 speculative flag 并明确验证范围
{
  match: { hw: "h100", strategy: "low-latency" },
  nnodes: 1,
  verified: true,
  // 仅当多模态传输走 auto 且 HiCache 关闭时，该组合才算已验证
  verificationStatus: (s) =>
    s.mmTransport === "auto" && s.hicache === "off"
      ? "verified"
      : "unverified",
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp-size 8",
    "--ep-size 8",
    "--mem-fraction-static 0.70", // 由 0.75 下调，避免 H100 96GB 显存预算下的 OOM 风险
    "--dsa-prefill-backend tilelang",
    "--dsa-decode-backend tilelang",
    "--kv-cache-dtype bfloat16",
    "--moe-runner-backend deep_gemm",
    "--disable-shared-experts-fusion",
    // 原 NEXTN 会被 SpeculativeAlgorithm.from_string 拒绝，MTP 目前统一由 EAGLE 承载
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 5",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 6",
    "--speculative-adaptive",
    "--reasoning-parser glm45",
    "--tool-call-parser glm47",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

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

基准数据表：为 h100/h200/b200/b300 八个空格 match 条目补录完整 GSM8K 精度数据、sglang_version 和复现说明，是 PR body 中评测结果的落地载体。

```
// 新增的 GSM8K 精度条目：记录复现命令、测量次数与口径限制
```

```
{
  match: { hw: "h100", strategy: "high-throughput" },
  sglang_version: "f040cc72e6",
  // 推荐 selection 的精度，4 个 selection 测量范围 97.27 - 97.50%
  accuracy: { gsm8k_pct: 97.50 },
  notes:
    "Full GSM8K (all 1,319 problems) on 8x H100 (TP8/EP8) with zai-org/GLM-5.3-Flash at
    f040cc72e6: 97.50% for the recommended selection; 97.27-97.50% across all 4 measured
    selections. Run with `sgl-eval run gsm8k --base-url http://localhost:30000/v1 --num-threads
    32 --max-tokens 32768`; gsm8k's registered default leaves thinking off, so these are non-
    thinking numbers and are not directly comparable to the GB300 rows above. Accuracy only,
    no speed measurement.",
},
```

评论区精华

本 PR 没有实质性的 review 交锋：唯一审批来自 wisclmy0611，状态为 APPROVED 且未留文字评论；PR 评论区仅有 Mintlify bot 的预览部署通知。技术决策（NEXTN 折叠为 EAGLE、H100 下调 mem-fraction-static）均以 commit message 形式记录，未展开讨论。

风险与影响

风险：

- 参数残留：仓库内其他模型 cookbook 页面可能仍有 NEXTN 写法，本 PR 只覆盖 GLM-5.3-Flash，需全仓库排查。
- 显存配置影响：H100 的 mem-fraction-static 从 0.75 降到 0.70 会减少 KV 缓存可用空间，长上下文或高并发用户可能有吞吐影响；旧文档复制 0.75 到 H100 可能 OOM。
- 基准口径差异：新记录是非 thinking、accuracy-only 数据，与 GB300 行不直接可比，跨行对比可能误读。
- verified 语义边界：34 个已验证组合不含 HiCache L3（Mooncake）与 GB200 / GB300 lane，这些组合上部署不能视为已验证。

影响：

- 用户侧：修复后的命令可直接复制运行，避免启动失败；各平台 GSM8K 精度数据为策略选择提供量化参考。
- 团队侧：cookbook 验证状态从 in-progress 升级为 verified，数据覆盖从仅 GB300 扩展到四种 NVIDIA 平台，文档可信度提升。
- 影响范围仅限文档，无运行时行为变更，影响程度中低。

关联脉络

- PR #36611 同为 cookbook 验证修正（Qwen3.8 Flash Next H200 MTP verify 与 BF16 SSM state），与本 PR 构成『cookbook 验证补全』的持续演进线。
- PR #34053 修正权重驻留内存与 KV sizing，本 PR 对 H100 / H200 显存占比的调整属于同一『内存与容量正确性』主题。

- PR #36396 在 MI30x 上补充 DSV4-Flash FP8 accuracy coverage，与本 PR 用完整 GSM8K 作为 verified 依据的方法论一致，说明团队正在把 accuracy gate 作为 cookbook 验证的标准手段。