

PR #36513 完整报告

sgl-project/sglang

GLM-5.3-Flash cookbook: FP8 KV + TRT-LLM benchmark cards and Blackwell default

合并时间: 2026-08-26 22:39

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36513>

PR 分析报告: GLM-5.3-Flash cookbook FP8 KV + TRT-LLM 基准卡与 Blackwell 默认配置

执行摘要

本 PR 为 GLM-5.3-Flash cookbook 新增了 FP8 KV cache + TRT-LLM DSA 组合的基准性能数据，并将其设为 Blackwell 平台（GB300 等）的默认 KV/DSA 配对。核心改动是扩展了基准查找逻辑 `findBenchmark`，使其支持基于附加维度（`kvDsaPair`）的匹配，并支持最具体条目优先。这解决了原 cookbook 中该选项无数据、且显示错误数据的问题。

功能与动机

原 cookbook (#36440) 提供 FP8 KV + TRT-LLM DSA 作为 Blackwell 的叠加选项，但基准卡片仅按硬件和策略键控，导致选择该选项后仍显示 BF16 + TileLang 的旧数据，且该选项没有任何测量数据（来自 PR body）。实测表明 FP8 KV + TRT-LLM 在所有行均更快，且 KV 容量约 1.8 倍、质量相当，因此本 PR 将其设为默认。

实现拆解

1. 扩展基准匹配逻辑：在 `docs/src/snippets/_deployment.jsx` 中，将 `findBenchmark` 从只能按 `DIMENSIONS` 精确匹配，改为支持任意 `match` 键，并采用“最具体匹配优先”策略。
2. 新增基准数据：在 `docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx` 中新增 4 条 FP8 + TRT-LLM 记录。
3. 调整默认与验证状态：在 `docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx` 中，GB300 的默认 `flags` 改为 FP8 + TRT-LLM，并扩展 `verificationStatus` 以认可该组合。
4. 测试与校验：PR 提到引擎 `harness` 和组合命令断言均通过，但未包含测试文件变更，可能为手动验证。

`docs/src/snippets/_deployment.jsx`

修改了基准查找逻辑 `findBenchmark`，使其支持基于 `overlay` 维度的匹配，这是本次变更的核心逻辑。

```
// findBenchmark 现在支持按 overlay 维度（如 kvDsaPair）匹配：
// 当一个条目的 match 中所有键都与当前选择相等时，该条目适用；
// 多个匹配时，声明键最多的条目（最具体）优先，
// 因此原有的 hwxstrategy 条目成为回退选项。
const findBenchmark = (list, sel) => {
```

```
// 过滤出所有 match 中每个键都与当前选择相等的条目
const hits = (list || []).filter((b) =>
  Object.entries(b.match || {}).every(([k, v]) => sel[k] === v));
// 按 match 声明的键数量降序排序，取最具体的条目
return hits.sort((a, b) =>
  Object.keys(b.match).length - Object.keys(a.match).length)[0] || null;
};
```

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

新增了 FP8 KV + TRT-LLM DSA 的基准数据，这些数据将用于展示特定组合的性能。

```
// 新增的基准条目示例：
// 通过 match 中的 kvDsaPair 区分同一硬件和策略下的不同 KV/DSA 组合。
{
  // 匹配 GB300 低延迟且 kvDsaPair 为 fp8-trtllm 的选择
  match: { hw: "gb300", strategy: "low-latency", kvDsaPair: "fp8-trtllm" },
  slang_version: "f13cb6f6a7",
  latencyPercentile: "Mean",
  speed: [
    {
      workload: { dataset: "random", isl: 1024, osl: 256,
        max_concurrency: 16, num_prompts: 80 },
      ttft_ms: 583.36,
      tpot_ms: 6.21,
      tokens_per_sec_per_gpu: 2357.1,
    },
  ],
}
```

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

调整了 GB300 的默认配置和验证状态，使其对 FP8 + TRT-LLM 组合显示为已验证。

```
// GB300 低延迟 cell 的验证状态：
// 当选择为推荐配置或满足特定条件（kvDsaPair 为 fp8-trtllm 且 mmTransport 为 auto、hcache 为 off）时，
// 标记为已验证，否则未验证。
verificationStatus: (s) =>
  config.isRecommendedSelection(s) ||
  (s.kvDsaPair === "fp8-trtllm" && s.mmTransport === "auto" && s.hcache === "off")
    ? "verified"
    : "unverified",

// 默认 flags 示例：切换 DSA 后端为 trtllm，KV cache dtype 为 fp8_e4m3
flags: [
  "--model-path {{MODEL_NAME}}",
  "--tp-size 4",
  "--ep-size 4",
  "--dsa-prefill-backend trtllm",
  "--dsa-decode-backend trtllm",
```

```
"--kv-cache-dtype fp8_e4m3",  
"--moe-runner-backend deep_gemm",  
// ... 其他参数  
]
```

评论区精华

本 PR 没有 review 评论，仅有一位审核者 APPROVED，无讨论内容。

风险与影响

- 默认配置变更：Blackwell 默认 KV dtype 从 BF16 改为 FP8，可能导致现有用户部署行为变化，需注意。
- 匹配逻辑变更：findBenchmark 匹配规则的改变可能影响其他基准条目的选取，需验证现有配置。
- 缺少测试：未发现对应的自动化测试，存在回归风险。

关联脉络

本 PR 是 #36440 GLM-5.3-Flash cookbook 的后续增强，体现了为特定硬件 / 软件组合提供精细基准数据的趋势。