PR #28062 完整报告

sgl-project/sglang

docs(minimax-m3): warm-steady-state benchmark numbers

合并时间：2026-06-12 23:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28062

执行摘要

此 PR 修正了 MiniMax-M3 cookbook 文档中的基准测试数据，将 B200 和 H200 的吞吐量从冷启动第一轮（由于预热慢约 2 倍）更新为 3 轮测试后的热稳态值。B200 的 tokens_per_sec_per_gpu 从 124 提升至 249，H200 从 105 提升至 116。同时补充了 GSM8K 准确率的稳定性说明及 B200 MSA 路径在负载下的准确率漂移问题。

功能与动机

PR body 明确指出："The MiniMax-M3 cookbook (merged in #28060) reported B200/H200 bench_serving throughput from the cold-start first run (cuda-graph capture + JIT warmup, ~2x slow). This corrects them to the warm steady-state from a 3-run sweep." 即原文档中的性能数据来自冷启动，包含额外的编译和预热开销，不能反映服务稳定后的真实吞吐。

实现拆解

1. 更新文件头部全局注释：将"run-1; a 3-run mean is pending"改为"warm steady-state from a 3-run sweep (the cold-start first run, ~2x slower, is excluded)"，并补充 GSM8K 的 3 轮结果详情——H200 稳定在 97.04%（std=0.0），B200 的 fresh-server 值在持续负载下会漂移。
2. 修改 B200 速度条目：将 ttft_ms 从 2410 改为 749，tpot_ms 从 148.4 改为 61.5，tokens_per_sec_per_gpu 从 124 改为 249。注释更新为"warm steady-state (3-run, cold-start run-1 excluded)"。
3. 修改 H200 速度条目：将 ttft_ms 从 1068 改为 1054，tpot_ms 从 78.0 改为 70.8，tokens_per_sec_per_gpu 从 105 改为 116。注释同步更新。
4. 更新 B200 GSM8K 注释：添加 NOTE 说明负载下准确率漂移现象（94.4→89.2→86.2），并强调这是服务问题，非模型准确率问题。
5. 更新 H200 GSM8K 注释：明确指明 3 轮测试均稳定在 97.04%（std=0.0）。

docs_new/src/snippets/configs/MiniMaxAI/minimax-m3-benchmarks.jsx

唯一修改的文件，包含 MiniMax-M3 的所有基准测试数据。更新了 B200 和 H200 的吞吐量、延迟和 GSM8K 准确率数据及注释。

```
// MiniMax-M3 per-cell benchmark numbers, keyed by the same `match` tuple as
// minimax-m3.jsx cells. See _deployment.jsx for the speed/accuracy schema.
//
```

```
// SPEED — bench_serving --flush-cache, random isl2048/osl256, max_concurrency 64,
// CUDA graph on. B200 (tp4, MXFP8, MSA fmha_sm100 path) and H200 (tp8, bf16,
// built-in Triton sparse) are measured on PR #27944 — warm steady-state from a
// 3-run sweep (the cold-start first run, ~2x slower, is excluded).
//
// GSM8K — unified on a SINGLE harness: sgl-eval `run gsm8k`, full 1319-questions.
// 3-run results: H200 is stable at 97.04% (std 0.0); B200's fresh-server 94.4%
// (= greedy) drifts down over sustained runs interleaved with bench (an
// MSA-under-load serving issue under investigation), so it reports the
// fresh-server value, not the drifted mean.
export const benchmarks = [
  {
    match: { hw: "b200", variant: "default", quant: "mxfp8", strategy: "balanced", nodes: "single" },
    sglang_version: "PR #27944",
    speed: [
      // bench_serving --flush-cache, MSA path; warm steady-state (3-run, cold-start run-1
      excluded).
      { workload: { dataset: "random", isl: 2048, osl: 256, max_concurrency: 64, num_prompts:
      128 },
        ttft_ms: 749, tpot_ms: 61.5, tokens_per_sec_per_gpu: 249 },
    ],
    accuracy: { gsm8k_pct: 94.4 }, // fresh-server 94.4% (greedy 94.16%; --no-thinking 88.6%).
    // NOTE: 3 sustained runs interleaved with bench drifted 94.4->89.2->86.2 —
    // an MSA-under-load serving issue (under investigation), not the model accuracy.
  },
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    sglang_version: "PR #27944",
    speed: [
      { workload: { dataset: "random", isl: 2048, osl: 256, max_concurrency: 64, num_prompts:
      128 },
        ttft_ms: 1054, tpot_ms: 70.8, tokens_per_sec_per_gpu: 116 },
    ],
    accuracy: { gsm8k_pct: 97.0 }, // stable 97.04% across all 3 runs (std 0.0)
  },
  // ... other platforms remain unchanged
];
```

评论区精华

PR 无 review 评论，审核人直接批准。无公开讨论。

风险与影响

纯文档变更，无代码逻辑或配置变动。无回归、性能、安全或兼容性风险。用户将看到更准确的性能数据，对选型决策有正面作用。B200 MSA 路径在负载下的准确率漂移说明增加了透明度，但该问题本身值得后续工程跟进。

关联脉络

关联 PR #28060：此 PR 是该 cookbook 文档的后续修正，将冷启动数据替换为热稳态数据。

- 关联 PR #27944：基准测试数据来源于该 PR（PR body 明确提及 "Docs-only; numbers measured on #27944"），但目前该 PR 的具体内容未知。