

# PR #28664 完整报告

sgl-project/sglang

[Cookbook] Laguna-M.1: enable FP8 on Blackwell + drop provisional AIME numbers

合并时间: 2026-06-19 00:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28664>

## 执行摘要

本 PR 更新了 Laguna-M.1 的 cookbook 配置与基准数据，在 Blackwell 平台上启用 FP8 量化（通过临时使用 Triton 后端避免 DeepGEMM 的 UE8M0 兼容问题），并移除了受截断偏差影响的 AIME 25 精度数据。变更仅涉及文档和基准配置文件，无运行时代码改动。

## 功能与动机

原 cookbook 页面 (#28661) 将 FP8 限制在 Hopper，因为 Blackwell 上 compressed-tensors block-FP8 权重缩放因子不是 UE8M0 格式，导致 DeepGEMM 产生 NaN。PR #28662 正在修复根本原因，本 PR 先通过 `--fp8-gemm-backend triton` 作为临时工作区，使用正确的 Triton 后端（性能慢约 19%）。同时，AIME 25 数据因 `max-tokens` 截断导致 ~33-37% 的样本分数为 0，暂移除以避免误导，直到无截断数据可用。

## 实现拆解

1. 修改 `laguna-m1.jsx` - 更新硬件注释，明确 FP8 在 Blackwell 上需加 `--fp8-gemm-backend triton`，并链接到 #28662。- 移除 `accuracy` 命令段中的 `aime25_pct` 条目，只保留 `gsm8k_pct`。- 调整 `defaultAccuracy` 和 `accuracyLabels`，移除 AIME 25 相关键。
2. 修改 `laguna-m1-benchmarks.jsx` - 添加 B200 FP8 基准条目，实测 GSM8K 为 93.78%（与 H200 FP8 的 93.25% 在噪声范围内一致）。- 移除所有 `aime25_pct` 字段，注释说明精度轴仅 GSM8K。- 新增 B300、GB200、GB300 的 FP8 待定桩。
3. 修改 `Laguna-M.1.mdx` - 更新 FP8 提示段落，明确 Blackwell 上需使用 `--fp8-gemm-backend triton`，并引用 #28662。

## `docs_new/src/snippets/configs/poolside/laguna-m1.jsx`

核心配置文件，定义硬件支持、量化组合、基准命令和精度标签。本 PR 扩展了 Blackwell FP8 支持注释，移除了 AIME 25 相关配置。

```
// 来自 laguna-m1.jsx，展示关键配置变更
// Hardware: H200 (Hopper) + B200/B300/GB200/GB300 (Blackwell).
// - BF16 runs everywhere.
// - FP8 runs everywhere. On Blackwell (sm_100) the compressed-tensors block-FP8 weight
scales
// aren't UE8M0-packed, so the default DeepGEMM path produces garbage → the Blackwell
FP8 cells
// add `--fp8-gemm-backend triton` (correct, ~19% slower than DeepGEMM). Temporary until
```

```

the
// ue8m0-requant fix (PR #28662) lands; H200 FP8 (Hopper) is unaffected and needs no flag.
// - NVFP4 is Blackwell-only.
// Cells: H200x{BF16,FP8}; each Blackwellx{BF16,FP8,NVFP4}.

```

```

export const config = {
  modelName: "Laguna-M.1",
  supportedHardware: ["h200", "b200", "b300", "gb200", "gb300"],
  variants: [ { id: "default", label: "Default" } ],
  quantizations: [
    { id: "bf16", label: "BF16" },
    { id: "fp8", label: "FP8" },
    { id: "nvfp4", label: "NVFP4" },
  ],
  strategies: [ { id: "balanced", label: "Balanced" } ],
  nodesOptions: [ { id: "single", label: "Single Node" } ],
  modelNames: {
    "defaultlbf16": "poolside/Laguna-M.1",
    "defaultlfp8": "poolside/Laguna-M.1-FP8",
    "defaultlnvfp4": "poolside/Laguna-M.1-NVFP4",
  },
  // accuracy 部分: 只保留 GSM8K, 移除 AIME 25
  defaultAccuracy: {
    default: { gsm8k_pct: null },
  },
  accuracyLabels: [
    ["gsm8k_pct", "GSM8K", "%"],
  ],
};

```

## docs\_new/src/snippets/configs/poolside/laguna-m1-benchmarks.jsx

基准数据文件，包含各硬件 + 量化组合的实测精度和性能数据。本 PR 添加了 B200 FP8 条目，移除了所有 AIME 25 数据。

```

// 来自 laguna-m1-benchmarks.jsx, 展示新增的 B200 FP8 条目
// 所有数据均为实际测量值, 精度轴仅 GSM8K (AIME 25 将在无截断数据时重新加入)
export const benchmarks = [
  // ===== H200 — BF16 / FP8 =====
  { match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    verified: true, sglang_version: "main @ 3f668733 (#28400 + #28604)", accuracy: { gsm8k_pct: 93.02 } },
  { match: { hw: "h200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" },
    verified: true, sglang_version: "main @ 3f668733 (#28400 + #28604 + g_proj FP8 fix #28649)",
    accuracy: { gsm8k_pct: 93.25 } },
  // ===== B200 — BF16 / FP8 / NVFP4 =====
  { match: { hw: "b200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    verified: true, sglang_version: "PR #28400 + #28604", accuracy: { gsm8k_pct: 91.88 } },
  { // 🔄 REAL — B200 FP8, with --fp8-gemm-backend triton
    match: { hw: "b200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" },

```

```
verified: true,  
sglang_version: "main + #28649 + --fp8-gemm-backend triton (DeepGEMM UE8M0  
workaround; fix = PR #28662)",  
accuracy: { gsm8k_pct: 93.78 } },  
{ match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },  
  verified: true, sglang_version: "PR #28400 + #28604", accuracy: { gsm8k_pct: 89.38 } },  
// ===== 其他 Blackwell 硬件为待定桩 =====  
{ match: { hw: "b300", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },  
{ match: { hw: "gb200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },  
{ match: { hw: "gb300", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },  
];
```

## 评论区精华

无实质性讨论。PR 获得直接批准 (reviewer: zijiaxia)。PR body 详细记录了技术背景和本地验证结果。

## 风险与影响

- 依赖 #28662: 若修复 PR 未能及时合入，本 PR 的临时 flag 可能长期作为唯一方案，且用户可能忽略 flag 导致空输出。
- 性能回退: Triton 后端比 DeepGEMM 慢约 19%，对延迟敏感的负载可能有影响。
- 文档过时: #28662 合并后需更新 cookbook 以移除 flag 说明。
- 数据完整性: AIME 25 的移除降低了覆盖的评估维度，但避免了对用户造成误导。

## 关联脉络

本 PR 是 Laguna-M.1 cookbook 系列的一部分：

- 28661 添加初始页面，FP8 仅限 Hopper。
- 28662 正在修复 DeepGEMM UE8M0 问题，本 PR 先行通过 Triton 临时启用。
- 28649 修复了 FP8 注意力门控量化配置，本 PR 的基准版本包含该修复。
- 后续 PR 将在 #28662 合并后移除 flag，并可能重新加入 AIME 25 数据。