

PR #28661 完整报告

sgl-project/sglang

Add Laguna-M.1 cookbook

合并时间: 2026-06-18 23:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28661>

执行摘要

本 PR 为 Poolside 的 Laguna-M.1 模型添加了完整的 SGLang 配置驱动 cookbook 页面，覆盖 H200/B200/B300/GB200/GB300 硬件平台，提供 BF16/FP8/NVFP4 量化的部署指南和验证基准。页面采用 JSX 配置与 MDX 内容分离的模式，支持自动生成部署命令。

功能与动机

根据 PR body: "Adds the config-driven SGLang Cookbook page for poolside/Laguna-M.1 under docs_new/." 目的是为社区提供官方部署指南，降低模型使用门槛，同时展示 SGLang 的配置驱动 cookbook 体系。该模型是 225B 参数 MoE (23B 激活)，支持 262K 上下文，Apache-2.0 协议。

实现拆解

1. 新增核心配置文件(laguna-m1.jsx): 定义模型名称、硬件支持列表、量化选项、部署策略、节点类型、模型名称映射、占位符变量、curl 示例、基准命令和默认精度格式。该文件被 _deployment.jsx 和 _playground.jsx 引擎消费。
2. 新增基准测试数据(laguna-m1-benchmarks.jsx): 为每个已验证的硬件 - 量化组合提供具体的性能数据 (TTFT、TPOT、吞吐量) 和精度分数 (GSM8K、AIME25)，并标注验证状态。未测量项留作 pending，杜绝虚构数字。
3. 新增主文档页面(Laguna-M.1.mdx): 包含模型介绍、安装指南 (Python/Docker 双标签)、通过导入配置和基准组件渲染的 Deployment 面板、Playground 实验区以及详细的推理和工具调用示例。
4. 更新导航配置(docs_new/docs.json): 在 Poolside 分组中添加新页面路径，确保页面在侧边栏和搜索引擎中出现。
5. 调整入口链接: 将 intro.mdx 中的 Poolside 卡片 href 从旧模型 Laguna-XS.2 改为新模型 Laguna-M.1，同时从 Laguna-XS.2.mdx 的 frontmatter 中移除 tag: NEW，将新标签转移至新页面。

docs_new/src/snippets/configs/poolside/laguna-m1-benchmarks.jsx

基准测试数据文件，提供了每个已验证组合的性能指标和精度数值，是用户评估部署的关键参考。

```
// laguna-m1-benchmarks.jsx — 每个已验证 cell 的基准数据
// 所有数字均为实测值，未测量项为 pending 桩 ("pending")
```

```

export const benchmarks = [
  // ===== H200 — BF16 =====
  {
    // 匹配条件（需与 laguna-m1.jsx 中的 cell 键完全一致）
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    verified: true, // 已验证
    sglang_version: "main @ 3f668733 (#28400 + #28604)", // 版本要求
    speed: [
      // 并发 1：低延迟场景
      { workload: { dataset: "random", isl: 4096, osl: 1024, max_concurrency: 1 },
        ttft_ms: 81.9, // 首 Token 时延中位数 (ms)
        tpot_ms: 8.91, // 单 Token 时延中位数 (ms)
        tokens_per_sec_per_gpu: 13.7 }, // 每 GPU 吞吐量 (tokens/s)
      // 并发 128：高吞吐场景
      { workload: { dataset: "random", isl: 4096, osl: 1024, max_concurrency: 128 },
        ttft_ms: 200.1,
        tpot_ms: 52.1,
        tokens_per_sec_per_gpu: 283 },
    ],
    accuracy: {
      gsm8k_pct: 93.02, // GSM8K 准确率 (%)
      aime25_pct: 53.33, // AIME25 准确率 (overall, 受 32k 截断限制, stop-only ~0.80)
    },
  },
  // ... 更多组合（H200 FP8、B200 BF16/NVFP4 等）
  {
    match: { hw: "b200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    verified: true,
    sglang_version: "PR #28400 + #28604",
    speed: [
      { workload: { dataset: "random", isl: 4096, osl: 1024, max_concurrency: 1 },
        ttft_ms: 108, tpot_ms: 9.0, tokens_per_sec_per_gpu: 13.6 },
      { workload: { dataset: "random", isl: 4096, osl: 1024, max_concurrency: 128 },
        ttft_ms: 170, tpot_ms: 43.3, tokens_per_sec_per_gpu: 331 },
    ],
    accuracy: { gsm8k_pct: 91.88, aime25_pct: 66.88 },
  },
  // 未验证的组合仅保留 match，界面显示 pending
  { match: { hw: "b300", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" } },
];

```

评论区精华

- gemini-code-assist[bot]在 laguna-m1.jsx:131 发现 dpAttn playground 控件的 labels 键名 "auto" 与 values 中的 null 不匹配，导致显示异常，建议改为 "null"。该评论未获回复，PR 直接合并，可能未被采纳。

风险与影响

- 风险：文档锁定特定 PR 版本和 Docker 镜像，如果后续主分支发生 breaking change，用户按文档操作可能遇到问题。部分平台组合未验证，用户需自行测试。
- 影响：为用户提供一键式部署体验，降低模型引入成本。推动 SGLang cookbook 体系扩展，增大模型覆盖面。

关联脉络

本 PR 依赖于三个核心功能 PR：

- #28400：softplus per-element 输出门控，实现模型原生支持。
- #28604：修复全局注意力滑动窗口问题，防止 M.1 在并发时崩溃。
- #28649：传递量化配置到注意力门控投影，修复 FP8 权重加载。

同时，本 PR 接替了之前 Laguna-XS.2 的入口位置，体现了 Poolside 模型系列在 SGLang 中的持续演进。