PR #31918 完整报告

sgl-project/sglang

[Cookbook] Add Laguna-S-2.1

合并时间：2026-07-22 01:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31918

执行摘要

本 PR 为 poolside 的 Laguna-S-2.1（118B MoE，8B active）添加完整的 SGLang cookbook 文档，包含多硬件 × 多量化的启动命令、经过验证的基准性能 / 准确率数据，并更新导航和首页入口。这是一个文档类变更，无代码风险，但基准数据依赖特定版本。

功能与动机

动机是为用户提供官方部署指南，降低使用门槛。PR body 列出了明确的验证条件：mint validate 通过、单元格渲染正确、URL-hash 持久化、NEW 标签仅新页面展示、首页卡片正确链接。

实现拆解

1. 配置层：laguna-s21.jsx 定义模型元数据、占位符、benchmark 命令、Playground 选项和 30+ 部署单元格。每个单元格匹配硬件 / 量化 / 策略 / 节点组合，包含环境变量和启动参数。
2. 基准数据层：laguna-s21-benchmarks.jsx 记录 H200/B300/GB300 在各量化下的速度（TTFT/TPOT/tokens_per_sec）和准确率（GSM8K/AIME25），均标注验证状态和 SGLang 版本。
3. 展示层：Laguna-S-2.1.mdx 导入上述数据，渲染 Deployment 和 Playground 交互组件，并包含安装步骤、模型简介、Configuration Tips。
4. 导航注册：docs.json 添加页面路径，intro.mdx 将 Poolside 卡片链接指向新模型。
5. 标签清理：从 Laguna-M.1.mdx 和 Laguna-XS-2.1.mdx 移除 tag: NEW。

docs_new/src/snippets/configs/poolside/laguna-s21-benchmarks.jsx

提供经过验证的速度和准确率基准数据，是用户选型的重要参考。

```
// laguna-s21-benchmarks.jsx — 基准测量数据
// 包含 H200/B300/GB300 在各量化下的速度（TTFT/TPOT）和准确率（GSM8K/AIME25）

export const benchmarks = [

  // —— H200 (8×H200, tp 8, sglang 0.5.15.post1) ——
  // 速度数据：random ISL=8192/OSL=1024，bench_serving 测量
  // tokens_per_sec_per_gpu 为每个 GPU 的吞吐

  // BF16 高吞吐：ttft 23.7s，tpot 256.8ms，吞吐 5264 tok/s/GPU
```

```
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "high-throughput", nodes:
    "single" },
    verified: true,
    sglang_version: "0.5.15.post1",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1024 },
        ttft_ms: 23742, tpot_ms: 256.8, tokens_per_sec_per_gpu: 5264 },
    ],
    // AIME25 为 null 因为 BF16 推理长度超 max_tokens=64000 导致无效
    accuracy: { gsm8k_pct: 93.18, aime25_pct: null },
  },

  // BF16 低延迟（concurrency=1 和 16 两个点）
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "low-latency", nodes: "single" }
    ,
    verified: true,
    sglang_version: "0.5.15.post1",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 102.3, tpot_ms: 3.48, tokens_per_sec_per_gpu: 305 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 97.7, tpot_ms: 6.45, tokens_per_sec_per_gpu: 2094 },
    ],
    accuracy: { gsm8k_pct: 93.33, aime25_pct: null },
  },

  // … 更多硬件 / 量化组合的条目
];
```

评论区精华

- 机器人：指出 NVFP4 验证状态在配置、基准、MDX 三处不一致，以及 Docker 标签 dev vs latest 矛盾。
- Jiminator：逐一回复已修复，NVFP4 在 stock 0.5.15 上验证通过，镜像统一为 latest。

风险与影响

风险：纯文档无运行时风险，但 benchmark 数值依赖特定版本（0.5.15.post1），版本升级后可能过时；命令参数如 --mem-fraction-static 若后端行为变化可能误导读者。影响：为用户提供直接可用的部署方案，提升 SGLang 对大型 MoE 模型的支持可见度。

关联脉络

本 PR 是 Poolside 系列 cookbook 的补充，与 Laguna-XS-2.1、Laguna-M.1 构成完整产品线。同期还添加了 Inkling 夜间测试（PR#31939），反映 SGLang 团队持续扩大模型覆盖范围。