

PR #27409 完整报告

sgl-project/sglang

add lfm2.5 to new cookbook.

合并时间: 2026-06-12 11:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27409>

执行摘要

本 PR 为 LiquidAI LFM2.5 模型系列添加了 config-driven 的 cookbook 页面，包含部署配置和基准性能数据，并更新了文档导航。用户可通过交互式面板生成部署命令和复现性能测试。

功能与动机

为 LiquidAI LFM2.5 模型（包括 8B MoE、1.2B Instruct/Thinking、350M、1.2B JP、VL 1.6B、VL 450M）提供官方部署指南和性能基准，降低用户使用门槛。所有部署配置经过验证，性能数据通过 SGLang 在 B200 上实际测量，确保准确性。

实现拆解

1. 新增部署配置片段: lfm2.5.jsx 定义 7 个变体参数、量化、策略、占位符和命令模板，覆盖三款 GPU。
2. 新增基准性能数据: lfm2.5-benchmarks.jsx 以 match 元组索引，提供每组合的 TTFT/TPOT/ 吞吐实测数据。
3. 编写 cookbook 页面: LFM2.5.mdx 集成 <Deployment> 组件，导入 config 和 benchmarks，生成交互式面板。
4. 更新导航: docs.json 新增 LiquidAI 分组; intro.mdx 添加卡片和 logo。
5. 添加 logo: liquidai.png 用于首页卡片。
6. Review 后调整: 移除 mode: wide，基准迁移到配置驱动，扩展 _deployment.jsx 的 ACCURACY_LABELS。

[docs_new/src/snippets/configs/LiquidAI/lfm2.5-benchmarks.jsx](#)

包含 LFM2.5 在 H100/H200/B200 上的详细性能基准数据，用于 benchmark 卡片

```
// 每个基准条目对应一个 (hw, variant, quant, strategy, nodes) 组合
export const benchmarks = [
  // ===== H100 =====
  {
    match: { hw: "h100", variant: "8b-a1b", quant: "bf16", strategy: "default", nodes: "single" },
    sglang_version: "0.0.0.dev1+g631db6c75",
    speed: [
      // 低并发 (latency) : 10 prompts, concurrency=1
      { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 1, num_prompts: 10 },
        ttft_ms: 287.24, tpot_ms: 2.4, tokens_per_sec_per_gpu: 325.11 },
```

```
// 高并发 (throughput): 1000 prompts, concurrency=100
{ workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 100, num_prompts:
1000 },
  ttft_ms: 171.72, tpot_ms: 11.87, tokens_per_sec_per_gpu: 7875.37 },
],
},
// 其余变体和硬件的条目结构相同，省略以保持简洁
];
```

评论区精华

"let's remove `mode: wide` here" — zijiaxia "Maybe remove the benchmark section and put all benchmark results into the `lfm2.5-benchmarks.jsx`?" — zijiaxia "is there any chance you can add the reproduce command for how you run those accuracy evaluation?" — zijiaxia 作者根据建议进行了相应调整，并扩展了 `_deployment.jsx` 以支持 LFM2.5 特有的评估键。

风险与影响

- 风险：仅文档变更，无运行时风险。基准数据可能随版本更新而过时，但附带版本号和复现命令。
- 影响：用户获得一站式 LFM2.5 部署指南；文档架构展示了 config-driven 新模式，便于后续扩展。

关联脉络

本 PR 是 LiquidAI 模型文档的首次引入。参考了 PR #27830 的 cookbook 样式，并在 review 中采纳了与之对齐的建议。后续可能继续完善其他模型的配置驱动文档。