PR #29321 完整报告

sgl-project/sglang

Add LFM2.5-230M to the LFM2.5 cookbook

合并时间：2026-06-26 08:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29321

执行摘要

PR #29321 在 SGLang 的 LFM2.5 cookbook 中集成最新发布的 230M 稠密模型，提供了完整的部署命令、精度基准和速度基准。变更仅涉及文档和配置数据，无运行时风险。

功能与动机

LiquidAI 公开了 LFM2.5-230M 模型，这是 LFM2.5 系列最紧凑的稠密变体。本 PR 将其加入现有的 cookbook，使用户可以像使用其他变体一样在 SGLang 中一键部署，并获取精度和速度参考数据。

实现拆解

1. 更新 lfm2.5.jsx：在 variants 列表添加 { id: "230m", label: "230M", subtitle: "dense · compact" }；在 modelNames 中加入 "230m|bf16" 到 HuggingFace 模型 ID 的映射；在 defaultAccuracy 中加入 "230m" 的 MMLU / GSM8K / GPQA 实测精度（38.45/31.84/27.78）；为 H100、H200、B200 各生成一个 verified 部署单元（TP=1、--tool-call-parser lfm2，B200 额外添加 --attention-backend trtllm_mha）。
2. 更新 lfm2.5-benchmarks.jsx：在 H100、H200、B200 的 benchmarks 数组中各插入一条 230M 的速度记录（latency c1: 546/561/1158 tok/s，throughput c100: 14280/17892/19206 tok/s）。
3. 更新 LFM2.5.mdx：在可用模型表格中添加 230M 行（230M 稠密、32K 上下文、适用场景为数据抽取与结构化输出）；将变体数量从 7 更新为 8；在推荐采样配置和工具调用列表中增加 230M；在 Base 检查点列表中也加入 230M。
4. 更新 generative_models.mdx：将 LFM2 型号列表从 (350M, 1.2B) 改为 (230M, 350M, 1.2B)。

docs_new/src/snippets/configs/LiquidAI/lfm2.5.jsx

cookbook 的部署配置核心文件，新增 230M 变体的 all 配置项（variants、modelNames、defaultAccuracy、verified cells），是本次变更的主要入口。

```
// Single `export const config` literal — no spreads/calls/IIFE (Mintlify re-evals at hydration).
// Cells are denormalized: no `--nnodes`/`--node-rank`/`--dist-init-addr`/`--host`/`--port` literals — engine injects them.

// LFM2.5 note: every variant runs on ONE GPU (TP=1), so the matrix is
// hw × variant only (single quant / strategy / nodes).
```

```
export const config = {
  modelName: "LFM2.5",
  supportedHardware: ["h100", "h200", "b200"],
  variants: [
    { id: "8b-a1b", label: "8B-A1B", subtitle: "8.3B MoE • reasoning" },
    { id: "instruct", label: "1.2B Instruct", subtitle: "1.17B dense" },
    { id: "thinking", label: "1.2B Thinking", subtitle: "1.17B • reasoning" },
    { id: "350m", label: "350M", subtitle: "dense" },
    { id: "230m", label: "230M", subtitle: "dense • compact" }, // <-- 新增变体
    { id: "jp", label: "1.2B JP", subtitle: "Japanese" },
    { id: "vl", label: "VL 1.6B", subtitle: "vision" },
    { id: "vl-450m", label: "VL 450M", subtitle: "vision • compact" },
  ],
  quantizations: [{ id: "bf16", label: "BF16" }],
  strategies: [{ id: "default", label: "Default" }],
  nodesOptions: [{ id: "single", label: "Single Node" }],
  modelNames: {
    "8b-a1b|bf16": "LiquidAI/LFM2.5-8B-A1B",
    "instruct|bf16": "LiquidAI/LFM2.5-1.2B-Instruct",
    "thinking|bf16": "LiquidAI/LFM2.5-1.2B-Thinking",
    "350m|bf16": "LiquidAI/LFM2.5-350M",
    "230m|bf16": "LiquidAI/LFM2.5-230M", // <-- 新增模型 ID
    "jp|bf16": "LiquidAI/LFM2.5-1.2B-JP-202606",
    "vl|bf16": "LiquidAI/LFM2.5-VL-1.6B",
    "vl-450m|bf16": "LiquidAI/LFM2.5-VL-450M",
  },
  defaultAccuracy: {
    "8b-a1b": { mmlu_pct: 76.61, gsm8k_pct: 91.96, gpqa_pct: 52.27, aime25_pct: 45.21 },
    thinking: { mmlu_pct: 63.2, gsm8k_pct: 86.35, gpqa_pct: 39.08, aime25_pct: 27.08 },
    instruct: { mmlu_pct: 60.33, gsm8k_pct: 75.13, gpqa_pct: 34.41, aime25_pct: 9.58 },
    "350m": { mmlu_pct: 40.69, gsm8k_pct: 30.63, gpqa_pct: 28.35 },
    "230m": { mmlu_pct: 38.45, gsm8k_pct: 31.84, gpqa_pct: 27.78 }, // <-- 新增精度数据
    vl: { mmmu_pct: 39.12 },
    "vl-450m": { mmmu_pct: 30.56 },
  },
  // ...（其余配置如 placeholders、curl、benchmarkCommands、dockerImages 等保持不变）
};
```

docs_new/src/snippets/configs/LiquidAI/lfm2.5-benchmarks.jsx

基准测试数据文件，为 230M 在三个硬件上提供实测速度数据，方便用户评估性能。

```
// benchmarks 数组中的新增条目示例（H100）：
{
  match: { hw: "h100", variant: "230m", quant: "bf16", strategy: "default", nodes: "single" },
  sglang_version: "0.0.0.dev1+g631db6c75",
  speed: [
    { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 1, num_prompts: 10 }
    ,
      ttft_ms: 23.74, tpot_ms: 1.77, tokens_per_sec_per_gpu: 546.07 },
```

```
      { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 100, num_prompts:
      1000 },
        ttft_ms: 1128.14, tpot_ms: 4.54, tokens_per_sec_per_gpu: 14280.32 },
    ],
  },
  // 类似条目也为 H200 和 B200 添加，分别标注 ttft_ms / tpot_ms / throughput。
```

评论区精华

无 review 讨论，PR 直接被批准。

风险与影响

无风险。影响限于 cookbook 页面更新，用户可查阅 230M 的部署与基准信息。

关联脉络

本 PR 是对 #27409（首次添加 LFM2.5 家族）和 #28072（精度基准重构）的后续扩展，延续了相同的配置结构和基准采集方法。