

PR #30320 完整报告

sgl-project/sglang

[Doc] Add LongCat 2.0 FP8 cookbook

合并时间: 2026-07-08 02:48

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30320>

执行摘要

本 PR 为 Meituan LongCat-2.0-FP8 模型添加了配置驱动的部署 cookbook 文档，基于 SGLang 文档模板，提供 Deployment 和 Playground 组件、基准数据和导航注册。B300 单节点部署已验证，其余硬件配置待验证。

功能与动机

为 Meituan LongCat-2.0-FP8 模型提供官方部署文档，遵循 SGLang 文档的 config-driven cookbook 模板，方便用户快速部署和调优。PR body 说明 'Regenerate the LongCat-2.0 FP8 cookbook with the config-driven Deployment / Playground template.'

实现拆解

1. 创建配置片段：在 docs_new/src/snippets/configs/meituan-longcat/longcat-2.0.jsx 中定义 config 对象，包含模型标识、支持的硬件、占位符、部署组合（cells）和 playground 功能。每个 cell 指定了匹配元组和启动 flags。
2. 添加基准数据：在 longcat-2.0-benchmarks.jsx 中导出 benchmarks 数组，包含 B300 单节点的 GSM8K 准确率 95.89%。
3. 编写 cookbook 页面：cookbook/autoregressive/Meituan/LongCat-2.0.mdx 引入配置和基准组件，提供安装指南、部署命令生成器和 Playground。
4. 注册到文档导航：修改 docs_new/docs.json，在 autoregressive 分组下新增 Meituan 组，指向该页面。
5. 添加概览卡片：修改 cookbook/autoregressive/intro.mdx，添加 Meituan 品牌卡片，并上传 cards/logos/meituan.png 图片。

docs_new/src/snippets/configs/meituan-longcat/longcat-2.0.jsx

核心配置文件，定义了模型标识、部署组合（cells）和 Playground 功能，是 cookbook 的首要数据来源。

```
// LongCat-2.0 部署配置对象（部分字段）
export const config = {
  modelName: "LongCat-2.0",
  // 支持的硬件标识列表
  supportedHardware: ["b300", "b200", "h200", "h20"],
  // 模型名映射，将 variantIquant 组合映射到 HuggingFace 模型名
  modelNames: {
```

```

    "defaultIfp8": "meituan-longcat/LongCat-2.0-FP8",
  },
  // 占位符定义，用户需在命令面板中填写
  placeholders: {
    HOST_IP: { target: "command", label: "Bind host", default: "0.0.0.0" },
    PORT: { target: "command", label: "Bind port", default: "30000" },
  },
  // 已验证的部署组合列表
  cells: [
    {
      match: { hw: "b300", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" },
      verified: true,
      flags: [
        "--trust-remote-code",
        "--model-path {{MODEL_NAME}}",
        "--tp 8",
        "--ep 8",
        "--kv-cache-dtype bfloat16",
      ],
    },
  ],
  // Playground 功能：允许用户调整高级选项
  playgroundFeatures: {
    moe: {
      backend: {
        options: [
          { id: null, label: "Inherited" },
          { id: "deepEP", label: "DeepEP", flags: ["--moe-a2a-backend deepEP"] },
        ],
      },
    },
  },
};

```

评论区精华

审核者 zijiexia 批准但指出两个后续事项：

- MTP 文档缺失：LongCat-2.0 有三阶段 MTP 模块但 cookbook 未提及 speculative decoding。
- 硬件验证不足：需要验证 H200/B200/H20 上的部署配置。这些尚未在本次 PR 中解决，已被标识为待处理。

风险与影响

- 配置风险：flags 硬编码在配置中（如 `--nsa-prefill-backend fa3`），不同硬件可能不兼容，用户若未注意 `verified` 标志可能误用。
- 验证不全：仅 B300 经过端到端验证，其他硬件配置可能存在问题。
- 维护成本：新增 cookbook 页面需要后续跟随模型更新。

- 影响：对用户是正向，降低了部署门槛；对团队是新增的文档维护任务。

关联脉络

本 PR 与 #30275 (Support LongCat 2.0 FP8) 直接关联，后者提供了 LongCat-2.0-FP8 模型的推理核心支持（包括 moe、attention 等），是本文档的前提。此外，PR #30372 (FlexKV 回归修复) 和 #30274 (DSA 优化) 虽不直接相关，但同属近期活跃的推理优化主线。