

PR #30201 完整报告

sgl-project/sglang

cookbook: add Hunyuan 3 (Hy3) Day-0 page

合并时间: 2026-07-06 13:30

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30201>

执行摘要

PR #30201 为 Tencent Hy3 (Hunyuan 3) MoE 模型新增了 Day-0 cookbook 文档页面，采用配置驱动模板实现多硬件、多量化、多策略的部署命令生成与基准展示。同步更新了站点导航和首页入口，并依赖 PR #29920 的 parser 改进使配置同时兼容预览版和正式版 tokenizer。

功能与动机

Hy3 是腾讯最新发布的约 295B 总参数量 / 21B 激活参数的 MoE 模型，支持混合思考、原生工具调用和 256K 上下文。cookbook 的动机是让用户能快速、可靠地部署该模型，通过一份配置覆盖 H200、B200、B300、GB200、GB300 等多种硬件，以及低延迟（EAGLE MTP）、均衡（chunked prefill）等策略。PR body 明确指出这是 "Day-0 cookbook page for Hunyuan 3 (Hy3) on the config-driven template"。

实现拆解

1. 模型配置(hy3.jsx): 定义了完整的配置对象，包括支持硬件列表、变体 / 量化选项、策略、模型名称映射、命令占位符、基准测试命令等。所有字段遵循共享的 _deployment.jsx 和 _playground.jsx 约定。
2. 基准数据(hy3-benchmarks.jsx): 以数组形式提供每个硬件 / 配置组合的 GSM8K 准确率，H200 BF16 两行已验证（95.75% / 95.83%），其余留作未验证占位。
3. 主页面(Hy3.mdx): 通过 import 引用配置和基准组件，利用 <Deployment> 标签自动渲染硬件选择、命令生成和基准显示。页面还包含安装说明、硬件规格表格和参数解析器提示。
4. 导航更新(docs.json): 在 Tencent 分组下添加 Hy3 页面，位于 Preview 页面之前。
5. 入口与标签清理: 将 intro.mdx 中的 Tencent 卡片链接指向 Hy3，并移除 Preview 页面的 tag: NEW。

关键源码片段

[docs_new/src/snippets/configs/tencent/hy3.jsx](#)

核心配置文件，定义模型名称、硬件支持、变体、量化、策略、命令占位符、基准命令等，是页面动态生成的基础。

```
// hy3.jsx — Hy3 cookbook 配置，驱动 _deployment.jsx 和 _playground.jsx
// 一份配置同时支持所有硬件、量化、策略组合，无需重复编写页面。
```

```

export const config = {
  modelName: "Hy3",
  supportedHardware: ["h200", "b200", "b300", "gb200", "gb300"],

  // 变体与量化分离, BF16 和 FP8 归为量化层级
  variants: [{ id: "default", label: "Default" }],
  quantizations: [
    { id: "bf16", label: "BF16" },
    { id: "fp8", label: "FP8" },
  ],

  // 部署策略: 低延迟 (EAGLE MTP) vs 均衡 (chunked prefill)
  strategies: [
    { id: "low-latency", label: "Low-Latency" },
    { id: "balanced", label: "Balanced" },
  ],

  // 模型名映射根据变体 + 量化确定 HuggingFace ID
  modelNames: {
    "defaultlbf16": "tencent/Hy3",
    "defaultlfp8": "tencent/Hy3-FP8",
  },

  // 模板占位符, 被 _deployment.jsx 用于生成实际命令
  placeholders: {
    HOST_IP: { target: "command", label: "Bind host", default: "0.0.0.0" },
    PORT: { target: "command", label: "Bind port", default: "30000" },
    CURL_HOST: { target: "curl", label: "Server host", default: "localhost" },
    CURL_PORT: { target: "curl", label: "Server port", default: "30000" },
  },

  // 基准命令, 会被注入到 Deployment 卡的 Benchmark 模态框中
  benchmarkCommands: {
    speed: `python3 -m sglang.bench_serving --backend sglang --host {{CURL_HOST}} --port {{CURL_PORT}} --model {{MODEL_NAME}} ...`,
    accuracy: {
      gsm8k_pct: `sgl-eval run gsm8k --base-url http://{{CURL_HOST}}:{{CURL_PORT}}/v1 --num-threads 32`,
      aime26_pct: `sgl-eval run aime26 --base-url http://{{CURL_HOST}}:{{CURL_PORT}}/v1 --model {{MODEL_NAME}} ...`,
    },
  },

  // 多节点提示 (仅 GB200)
  multiNodeHints: {
    gb200: [
      "GLOO_SOCKET_IFNAME=<your-nic>",
      "NVSHMEM_ENABLE_NIC_PE_MAPPING=1",
      "NVSHMEM_HCA_LIST=<your-hca-list>",
    ],
  },
};

```

```

    ],
  },

  // Docker 镜像使用 dev 标签, 内含后缀感知 parser
  dockerImages: {
    h200: "lmsysorg/sglang:dev",
    b200: "lmsysorg/sglang:dev",
    b300: "lmsysorg/sglang:dev",
    gb200: "lmsysorg/sglang:dev",
    gb300: "lmsysorg/sglang:dev",
  },
};

```

docs_new/src/snippets/configs/tencent/hy3-benchmarks.jsx

提供各硬件 / 配置组合的 GSM8K 准确率数据, 用于 Deployment 卡中的 Benchmark 模态框展示。H200 BF16 单元格已验证, 其余留作占位。

```

// hy3-benchmarks.jsx — 各硬件配置的 GSM8K 准确率数据
// 键 match 与 hy3.jsx 中的下拉选项一一对应
// H200 BF16 两行已验证, 其余 (FP8 及 Blackwell) 为 Day-0 未验证占位
export const benchmarks = [
  // H200 BF16 已验证
  { match: { hw: "h200", variant: "default", quant: "bf16", strategy: "low-latency", nodes: "single" },
    , gsm8k_pct: 95.75 },
  { match: { hw: "h200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" },
    , gsm8k_pct: 95.83 },

  // BF16 其他硬件未验证, 留空对象
  { match: { hw: "b200", variant: "default", quant: "bf16", strategy: "low-latency", nodes: "single" }
  },
  { match: { hw: "b200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" } },
  { match: { hw: "b300", variant: "default", quant: "bf16", strategy: "low-latency", nodes: "single" }
  },
  { match: { hw: "b300", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" } },
  { match: { hw: "gb300", variant: "default", quant: "bf16", strategy: "low-latency", nodes:
    "single" } },
  { match: { hw: "gb300", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" }
  },
  { match: { hw: "gb200", variant: "default", quant: "bf16", strategy: "low-latency", nodes:
    "single" } },
  { match: { hw: "gb200", variant: "default", quant: "bf16", strategy: "balanced", nodes: "single" }
  },

  // FP8 所有单元格均未验证
  { match: { hw: "h200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" }
  },
  { match: { hw: "h200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },
  { match: { hw: "b200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" }
  },

```

```
{ match: { hw: "b200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },
{ match: { hw: "b300", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" } },
},
{ match: { hw: "b300", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },
{ match: { hw: "gb300", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" } },
},
{ match: { hw: "gb300", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },
{ match: { hw: "gb200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" } },
},
{ match: { hw: "gb200", variant: "default", quant: "fp8", strategy: "balanced", nodes: "single" } },
];
```

评论区精华

- JSX 表格: reviewer zijiexia 要求将 Markdown 表格改为 JSX 表格以改善渲染, 作者遵照修改。
- GB200 TP 缩放: zijiexia 质疑 GB200 的 TP 配置可能无法容纳模型, 作者解释 300B 模型可适应, 后随提交调整。
- 模型名称: zijiexia 确认模型仓库是 tencent/Hy3 而非 tencent/Hy3-preview, 作者已更正。
- 基准位置: zijiexia 建议将独立基准章节合并到 Deployment 卡内, 作者删除冗余 §4。

风险与影响

- 配置准确性: FP8 和 Blackwell 单元格均未验证, 用户在使用这些配置时可能遇到未预期的问题。
- Docker 镜像依赖: 当前使用 lmsysorg/sglang:dev 标签, 若正式发布后未及时切换至: latest, 用户可能使用不稳定版本。
- 硬件参数: GB200 的 TP 值虽经调整, 但实际部署前仍需进一步验证。
- 影响范围: 主要影响阅读文档的 Hy3 部署用户, 不涉及代码运行。

关联脉络

该 PR 依赖 PR #29920 提供的后缀感知 parser, 使得 `--reasoning-parser auto` `--tool-call-parser auto` 能自动适配不同版本的 Hy3 tokenizer。此前已有多个 cookbook 页面 (如 MiniMax、GLM) 采用相同的配置模板模式, 本 PR 延续了这一架构, 并进一步规范了多硬件多策略的配置组织方式。