PR #23532 完整报告

sgl-project/sglang

docs: add Hunyuan 3 Preview cookbook

合并时间：2026-04-23 17:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23532

执行摘要

本 PR 为 Tencent Hunyuan 3 Preview 模型添加了完整的部署文档 cookbook，包括模型介绍、硬件需求、部署命令交互式生成器以及性能基准，并更新了文档导航。这是纯文档变更，不涉及运行时代码，但依赖模型代码后续 upstream。

功能与动机

Hunyuan 3 Preview 是腾讯发布的第三代 MoE 语言模型，拥有约 276B 参数（20B 激活），支持混合推理、原生工具调用和 MTP 推测解码。PR body 指出模型代码尚未上游，文档先行可为用户提供即时的部署参考。

实现拆解

1. 编写文档主体 (Hunyuan3-Preview.mdx)

在 docs_new/cookbook/autoregressive/Tencent/ 下新增 527 行 Mintlify 格式文档，包含：

- 模型规格：MoE 架构、混合推理模式（reasoning_effort）、256K 上下文、MTP 模块。
- 安装方式：Docker 镜像表（含占位符 lmsysorg/sglang:hy3-preview）。
- 部署示例：基础部署、混合推理、非流式 / 流式工具调用。
- 基准数据：GSM8K 95.0%、MMLU 82.5%、工具调用准确率 100%。

2. 开发交互式部署命令生成器 (hunyuan3-preview-deployment.jsx)

新增 React 组件 Hunyuan3PreviewDeployment，用户可通过单选按钮选择硬件（H200/B200/B300/GB300）、推理解析器、工具调用解析器和推测解码，实时生成 sglang serve 命令。核心逻辑在 generateCommand 函数中：

- 根据硬件选择对应的 tp 值和 mem-fraction-static。
- 支持四种组合：--reasoning-parser hunyuan / --tool-call-parser hunyuan / --speculative-algorithm EAGLE 参数。
- 对 Blackwell 硬件（B200/B300/GB300）自动添加 --attention-backend trtllm_mha。

3. 更新导航配置 (docs.json)

在 Autoregressive Models 分组的末尾添加 Tencent 子分组，包含新文档页面。

docs_new/src/snippets/autoregressive/hunyuan3-preview-deployment.jsx

交互式部署命令生成器，用户通过单选按钮调整参数，实时生成 sglang serve 命令。

关键源码片段

docs_new/src/snippets/autoregressive/hunyuan3-preview-deployment.jsx

交互式部署命令生成器，用户通过单选按钮调整参数，实时生成 sglang serve 命令。

```
export const Hunyuan3PreviewDeployment = () => {
  // 定义可选的硬件平台及对应的 tp 值和显存占用
  const options = {
    hardware: {
      name: 'hardware',
      title: 'Hardware Platform',
      items: [
        { id: 'h200', label: 'H200', default: true },
        { id: 'b200', label: 'B200', default: false },
        { id: 'b300', label: 'B300', default: false },
        { id: 'gb300', label: 'GB300', default: false }
      ]
    },
    reasoning: {
      name: 'reasoning',
      title: 'Reasoning Parser',
      items: [
        { id: 'disabled', label: 'Disabled', default: false },
        { id: 'enabled', label: 'Enabled', default: true }
      ]
    },
    toolcall: {
      name: 'toolcall',
      title: 'Tool Call Parser',
      items: [
        { id: 'disabled', label: 'Disabled', default: false },
        { id: 'enabled', label: 'Enabled', default: true }
      ]
    },
    speculative: {
      name: 'speculative',
      title: 'Speculative Decoding (MTP)',
      items: [
        { id: 'disabled', label: 'Disabled', default: true },
        { id: 'enabled', label: 'Enabled', subtitle: 'Low Latency', default: false }
      ]
    }
  };

  // 每个硬件对应的配置：tp 和 mem_fraction
  const modelConfigs = {
    h200: { tp: 8, mem: 0.9 },
```

```
  b200: { tp: 8, mem: 0.9 },
  b300: { tp: 4, mem: 0.9 },
  gb300: { tp: 4, mem: 0.9 }
};

// ... ( 中间的状态管理和暗模式检测略 )

/**
 * 根据用户选择生成 sglang serve 命令
 * - 若启用推测解码，前置环境变量 SGLANG_ENABLE_SPEC_V2=1
 * - Blackwell 硬件自动追加 --attention-backend trtllm_mha
 */
const generateCommand = () => {
  const { hardware } = values;
  const isBlackwell = hardware === 'b200' || hardware === 'b300' || hardware === 'gb300';
  const hwConfig = modelConfigs[hardware];
  if (!hwConfig) return '# Configuration not available for the selected hardware.';

  const modelName = 'tencent/Hy3-preview';
  const tpValue = hwConfig.tp;
  const memFraction = hwConfig.mem;
  const enableSpec = values.speculative === 'enabled';

  let cmd = '';
  if (enableSpec) cmd += 'SGLANG_ENABLE_SPEC_V2=1 ';
  cmd += 'sglang serve \\\n';
  cmd += `  --model-path ${modelName}`;
  cmd += ` \
--tp ${tpValue}`;

  if (values.reasoning === 'enabled') cmd += ' \
--reasoning-parser hunyuan';
  if (values.toolcall === 'enabled') cmd += ' \
--tool-call-parser hunyuan';
  if (enableSpec) {
    cmd += ' \
--speculative-algorithm EAGLE';
    cmd += ' \
--speculative-num-steps 3';
    cmd += ' \
--speculative-eagle-topk 1';
    cmd += ' \
--speculative-num-draft-tokens 4';
  }
  cmd += ' \
--trust-remote-code';
  cmd += ` \
--mem-fraction-static ${memFraction}`;
  if (isBlackwell) cmd += ' \
```

```
    --attention-backend trtllm_mha';
      return cmd;
    };
```

评论区精华

无 review 讨论，审核者直接批准。

风险与影响

- 占位符风险：Docker tag 和 License 字段为占位符，需在模型代码发布后更新，否则用户可能使用无效 tag。
- 依赖未上游：模型代码、解析器和 MTP 加载器未合入，文档中的部署命令暂时无法使用。
- 硬件假设：交互生成器移除了 A100/H100，仅支持 80GB+ 显存 GPU，若未来提供更低精度版本需调整。

关联脉络

此 PR 是独立文档更新，与近期其他文档 PR（如弃用通知横幅 #23516）无直接关联，但共同完善了 docs_new/ 站点内容。模型代码对应的解析器 PR（如 hunyuan 工具调用解析器）预计将后续提交。