PR #24991 完整报告

sgl-project/sglang

[Docs] Update MiniCPM-V-4.6 documentation and deployment configuration

合并时间：2026-05-12 02:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24991

执行摘要

本 PR 更新了 MiniCPM-V-4.6 的文档和部署交互组件，新增 Thinking 变体支持，调整工具调用解析器，并移除预览状态标记。变更仅涉及文档和前端组件，无运行时影响。

功能与动机

MiniCPM-V-4.6 模型已在 HuggingFace 正式发布 Base 和 Thinking 两个变体，工具调用格式也升级为 Qwen3.5 XML 结构。为使用户能正确部署和使用新变体，需要同步更新部署配置文档，并清理之前占位的预览提示。

实现拆解

1. 新增 variant 选项：在部署交互组件 minicpm-v-4_6-deployment.jsx 的选项配置中新增 variant 字段，允许用户选择 base 或 thinking 变体。默认值为 base。
2. 动态模型路径：命令生成函数 generateCommand 根据 variant 值动态选择模型路径：openbmb/MiniCPM-V-4.6 或 openbmb/MiniCPM-V-4.6-Thinking。
3. 工具调用解析器更新：将工具调用解析器从 qwen 改为 qwen3_coder，以匹配新版本的工具调用 XML 格式。
4. 调整 Reasoning Parser 默认值：Reasoning Parser 的默认状态从 enabled 改为 disabled，因为 Thinking 变体本身已具备思考能力。
5. 更新 cookbook 文档：在 MiniCPM-V-4_6.mdx 中补充模型介绍、开源许可证（Apache 2.0）、基准测试结果，并移除预览期占位注释。

docs_new/src/snippets/autoregressive/minicpm-v-4_6-deployment.jsx

部署交互组件，新增 variant 选项，实现模型路径动态选择，是前端展示的核心变更。

```
// 选项配置：包含硬件、变体、推理、工具调用、Mamba 缓存等
const options = {
  hardware: {
    name: 'hardware',
    title: 'Hardware Platform',
    items: [
      { id: 'a100', label: 'A100', default: false },
      { id: 'h100', label: 'H100', default: false },
      { id: 'h200', label: 'H200', default: true },
      { id: 'b200', label: 'B200', default: false },
    ],
```

```
  },
  variant: { // 新增：Base / Thinking 变体选择
    name: 'variant',
    title: 'Variant',
    items: [
      { id: 'base', label: 'Base', subtitle: 'MiniCPM-V-4.6', default: true },
      { id: 'thinking', label: 'Thinking', subtitle: 'MiniCPM-V-4.6-Thinking', default: false },
    ],
  },
  reasoning: { name: 'reasoning', title: 'Reasoning Parser',
    items: [
      { id: 'enabled', label: 'enabled', default: false },
      { id: 'disabled', label: 'disabled', default: true },
    ],
  },
  toolcall: { name: 'toolcall', title: 'Tool Call Parser',
    items: [
      { id: 'enabled', label: 'enabled', default: false },
      { id: 'disabled', label: 'disabled', default: true },
    ],
  },
  mambaCache: { name: 'mambaCache', title: 'Mamba Radix Cache',
    items: [
      { id: 'v1', label: 'V1', default: false },
      { id: 'v2', label: 'V2', default: true },
    ],
  },
};

// 各硬件推荐的 tp 和 mem-fraction-static
const modelConfigs = {
  a100: { tp: 1, mem: 0.7 },
  h100: { tp: 1, mem: 0.7 },
  h200: { tp: 1, mem: 0.5 },
  b200: { tp: 1, mem: 0.4 },
};

// 核心命令生成函数
const generateCommand = (values) => {
  const { variant, hardware, reasoning, toolcall, mambaCache } = values;
  const hwConfig = modelConfigs[hardware];
  if (!hwConfig) return `# Error: Unknown hardware platform`;

  const { tp, mem } = hwConfig;
  const isBlackwell = hardware === 'b200';
  // 根据变体选择模型路径
  const modelPath = variant === 'thinking'
    ? 'openbmb/MiniCPM-V-4.6-Thinking'
    : 'openbmb/MiniCPM-V-4.6';
```

```
    let cmd = `sglang serve --model-path ${modelPath}`;
    if (tp > 1) cmd += ` \
    --tp ${tp}`;
    cmd += ` \
    --trust-remote-code`;
    cmd += ` \
    --dtype bfloat16`;
    if (isBlackwell) cmd += ` \
    --attention-backend trtllm_mha`;
    cmd += ` \
    --mem-fraction-static ${mem}`;
    if (reasoning === 'enabled') cmd += ` \
    --reasoning-parser qwen3`;
    if (toolcall === 'enabled') cmd += ` \
    --tool-call-parser qwen3_coder`; // 更新为 qwen3_coder
    if (mambaCache === 'v2') cmd += ` \
    --mamba-scheduler-strategy extra_buffer`;
    cmd += ` \
    --host 0.0.0.0 --port 30000`;

    return cmd;
  };
```

评论区精华

PR 无 review 评论，由 wisclmy0611 直接批准合并。

风险与影响

无运行时风险。文档和前端组件变更独立，不会影响已有服务。用户将获得更准确的部署指引。

关联脉络

该 PR 是 MiniCPM-V-4.6 模型集成工作的后续文档更新，与之相关的前期 PR 包括初始模型支持（未列出）等。没有发现其他直接关联的开放 PR。