

PR #27714 完整报告

sgl-project/sglang

[Docs] Add Kimi-K2.6 NVFP4 and update Kimi-K2.5 cookbook guidance

合并时间: 2026-06-11 01:24

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27714>

PR 分析报告：添加 Kimi-K2.6 NVFP4 部署指南并更新 K2.5 文档

1. 执行摘要

本 PR 为 Kimi-K2.6 模型新增了 NVFP4 量化部署指南，并同步更新了 Kimi-K2.5 文档。主要变更集中在两个 JSX 部署命令生成器和两个 MDX 文档上，通过动态选项和条件渲染处理了 Blackwell NVFP4 专属性和硬件切换时的状态一致性。用户可通过交互式生成器获得准确的部署命令，降低了配置错误概率。

2. 功能与动机

随着 NVIDIA Blackwell 架构发布 NVFP4 量化模型 [nvidia/Kimi-K2.6-NVFP4](#)，需要为 Kimi-K2.6 提供官方部署指导。同时修正 K2.5 文档中的过时内容：B200 已不再单独列出，[SGLANG_ENABLE_SPEC_V2=1](#) 环境变量已不再需要 (spec v2 默认启用)，且草稿模型名称从 [kimi-k2.5-eagle3](#) 更新为 [kimi-k2.5-eagle3-mla](#) 以匹配最新模型发布。

3. 实现拆解

1. 重构 Kimi-K2.6 部署命令生成器([kimi-k26-deployment.jsx](#)): 移除 B200/GB200 硬件选项，新增量化动态选项 (NVFP4 仅 Blackwell 可选)，新增推测解码选项 (带 condition 条件，仅非 AMD 硬件显示)，并添加 [useEffect](#) 监听硬件变化以重置失效选项，同时在 [generateCommand](#) 中增加双重检查。
2. 同步更新 Kimi-K2.5 部署命令生成器([kimi-k25-deployment.jsx](#)): 添加 GB300 硬件选项 (tp=4)，扩展 NVFP4 支持到 GB300，更新推测解码条件包含 GB300，移除所有 [SGLANG_ENABLE_SPEC_V2=1](#) 环境变量，并将草稿模型路径统一为 [lightseekorg/kimi-k2.5-eagle3-mla](#)。
3. 更新 Kimi-K2.6 MDX 文档：列出可用的 NVFP4 checkpoint，添加 NVFP4 部署说明 (命令、TP 配置)，新增推测解码小节，引用 NVIDIA 精度基准结果。
4. 更新 Kimi-K2.5 MDX 文档：硬件列表增加 GB300，更新所有命令示例 (移除 spec v2 环境变量，使用 EAGLE3 MLA 草稿模型，将 B200 引用改为 B300)，在 NVFP4 命令后添加 GB300 tp=4 说明。

[docs_new/src/snippets/autoregressive/kimi-k26-deployment.jsx](#)

核心变更文件，重构部署命令生成器，添加量化动态选项、推测解码条件渲染和硬件切换状态重置逻辑

```

export const KimiK26Deployment = () => {
  // Config mirrors sgl-cookbook src/components/autoregressive/KimiK26ConfigGenerator/index.
  js.
  //
  // INT4: H200/B300: tp=8, GB300/AMD: tp=4
  // NVFP4: B300: tp=8, GB300: tp=4
  const options = {
    hardware: {
      name: 'hardware',
      title: 'Hardware Platform',
      items: [
        { id: 'h200', label: 'H200', default: true },
        { id: 'b300', label: 'B300', default: false },
        { id: 'gb300', label: 'GB300', default: false },
        // AMD options omitted for brevity
      ],
    },
    quantization: {
      name: 'quantization',
      title: 'Quantization',
      getDynamicItems: (values) => {
        const hw = values.hardware;
        const isBlackwell = ['b300', 'gb300'].includes(hw);
        return [
          { id: 'int4', label: 'INT4', subtitle: 'Base checkpoint', default: !isBlackwell },
          { id: 'nvfp4', label: 'NVFP4', subtitle: 'Blackwell FP4', default: isBlackwell, disabled:
            !isBlackwell, disabledReason: !isBlackwell ? 'NVFP4 only on NVIDIA Blackwell' : '' },
        ];
      },
    },
    speculative: {
      name: 'speculative',
      title: 'Speculative Decoding',
      // 仅在非 AMD 硬件上显示此选项 (AMD 不支持 EAGLE3 MLA)
      condition: (values) => ![ 'mi300x', 'mi325x', 'mi350x', 'mi355x' ].includes(values.hardware),
      items: [
        { id: 'disabled', label: 'Disabled', default: true },
        { id: 'enabled', label: 'Enabled', default: false },
      ],
    },
  },
  // 其他选项 (reasoning, toolcall, dpattention) 省略
};

```

// 硬件切换时重置可能失效的选项

```

useEffect(() => {
  setValues((prev) => {
    const next = { ...prev };
    for (const [key, option] of Object.entries(options)) {
      if (typeof option.condition === 'function' && !option.condition(next)) {

```

```

// 当条件不满足时，将选项重置为默认可用值
const items = resolveItems(option, next);
const fallback = items.find((item) => item.default && !item.disabled) || items.find((item) => !item.disabled);
if (fallback) next[key] = fallback.id;
continue;
}
// 同时处理 getDynamicItems 变化（此处省略）
}
return next;
});
}, [values.hardware]);

const generateCommand = () => {
  const { hardware, quantization, speculative } = values;
  const isAMD = hardware.startsWith('mi');
  // 双重检查：即使选项 state 为 enabled，但硬件不满足时输出提示而非错误命令
  if (speculative === 'enabled' && isAMD) {
    return '# Speculative Decoding for Kimi-K2.6 is only supported on NVIDIA GPUs (H200/B300/GB300)';
  }
  // 继续生成命令 ...
};
// 其他逻辑 ...
};

```

5. 评论区精华

- 推测解码草稿模型选择：b8zhong 询问是否应使用 EAGLE3.1 checkpoint，mmangkad 回应 K2.5 尚无 EAGLE3.1，保持使用当前 MLA 版本。
- Speculative 条件渲染漏洞：zijiexia 指出 useEffect 未检查 option.condition，导致 AMD 上 speculative 仍可启用并生成错误命令。mmangkad 随后在 generateCommand 中添加额外守卫检查，并在第三次提交中修复了硬件切换时的状态残留问题。

6. 风险与影响

- 风险：文档与代码逻辑同步风险（未来硬件变化需同时更新 .jsx 和 .mdx）；硬件兼容性假设风险（NVFP4 仅 Blackwell）；缺少测试覆盖。
- 影响：用户获得更准确的部署命令，降低配置错误率；系统无运行时变更；团队需在后续硬件更新时维护这些生成器。

7. 关联脉络

本 PR 与 [#27668](#) (MiMo-V2.5-Pro 部署文档修复) 类似，都是对 cookbook 中部署命令生成器的修正和增强。两者共同反映了 SGLang 在支持多硬件和量化方案时对文档准确性的持续投入。未来 MoonshotAI Kimi 系列模型的文档维护可沿用本 PR 确立的条件渲染模式和状态重置策略。