PR #23540 完整报告

sgl-project/sglang

docs: split MI300X and MI325X options in GLM-5.1 generator

合并时间：2026-04-24 03:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23540

执行摘要

本 PR 将 GLM-5.1 部署交互组件中的 AMD 硬件选项从合并的“MI300X/MI325X”拆分为独立的“MI300X”和“MI325X”，确保两者使用相同的 BF16 配置，并重新排序了文档导航以突出最新版本。变更仅涉及文档站点 UI 组件和导航 JSON，无服务端代码或模型逻辑改动，风险极低。

功能与动机

PR Body 明确说明需要拆分 MI300X 和 MI325X 的硬件选项，让用户能够针对具体硬件进行选择，同时保持两者底层配置一致。此外，调整 GLM cookbook 页面排序，使较新模型（如 GLM-5.1）优先展示，提升文档导航的合理性和用户体验。

实现拆解

1. 硬件选择器拆分与 AMD 逻辑重构（docs_new/src/snippets/autoregressive/glm-51-deployment.jsx）：
   - 将硬件列表中的一项 { id: 'mi300x', label: 'MI300X/MI325X' } 拆分为 { id: 'mi300x', label: 'MI300X' } 和 { id: 'mi325x', label: 'MI325X' } 两项。
   - 将所有 AMD 检测条件（如 isAMD、投机解码条件、命令生成中的判断）从硬编码 === 比较改为 ['mi300x', 'mi325x', 'mi355x'].includes(hw)，提升了可读性和后续扩展性。
   - 在 modelConfigs 中为 mi325x 添加了与 mi300x 相同的配置项：{ bf16: { tp: 8, mem: 0.80 } }。

核心代码示例：
```
// 硬件选项从合并项拆分为独立项
items:[
// ...
{id:'mi300x',label:'MI300X',default:false},
{id:'mi325x',label:'MI325X',default:false},
{id:'mi355x',label:'MI355X',default:false}
]
// AMD 检测统一为 includes 写法
constisAMD=['mi300x','mi325x','mi355x'].includes(hw);
```

1. 导航排序调整（docs_new/docs.json）：
   - 将 GLM 组内页面顺序从版本升序（4.5, 4.6, 4.7, 4.7-Flash, 5, 5.1, Glyph, OCR, 4.5V, 4.6V）翻转为版本降序（5.1, 5, OCR, Glyph, 4.7, 4.7-Flash, 4.6, 4.6V, 4.5, 4.5V）。
2. 目录页链接更新（docs_new/cookbook/autoregressive/intro.mdx）：
   - 将 GLM 卡片的目标链接从 /cookbook/autoregressive/GLM/GLM-4.5 改为 /cookbook/autoregressive/GLM/GLM-5.1，使用户点击后直接跳转到最新版本页面。

docs_new/src/snippets/autoregressive/glm-51-deployment.jsx

核心变更文件，拆分硬件选项并重构 AMD 判断逻辑

关键源码片段

docs_new/src/snippets/autoregressive/glm-51-deployment.jsx

核心变更文件，拆分硬件选项并重构 AMD 判断逻辑

```
export const GLM51Deployment = () => {
  // 硬件选择器从合并选项拆分为 MI300X 和 MI325X 独立项
  const options = {
    hardware: {
      name: 'hardware',
      title: 'Hardware Platform',
      items: [
        { id: 'h200', label: 'H200', default: true },
        { id: 'b200', label: 'B200', default: false },
        { id: 'gb300', label: 'GB300', default: false },
        { id: 'h100', label: 'H100', default: false },
        { id: 'mi300x', label: 'MI300X', default: false }, // 拆分自合并项
        { id: 'mi325x', label: 'MI325X', default: false }, // 新增独立项
        { id: 'mi355x', label: 'MI355X', default: false }
      ]
    },
    // ... 其他配置保持不变
    speculative: {
      name: 'speculative',
      title: 'Speculative Decoding',
      // 条件改为 includes 以包含 mi325x
      condition: (values) => !['mi300x', 'mi325x', 'mi355x'].includes(values.hardware),
      items: [
        { id: 'disabled', label: 'Disabled', default: false },
        { id: 'enabled', label: 'Enabled', default: true }
      ]
    }
  };

  // 为 MI325X 添加与 MI300X 相同的模型配置（BF16, tp=8, mem=0.80）
  const modelConfigs = {
    // ... 其他硬件配置
    mi300x: { bf16: { tp: 8, mem: 0.80 } },
    mi325x: { bf16: { tp: 8, mem: 0.80 } }, // 新增项
    mi355x: { bf16: { tp: 8, mem: 0.80 } }
  };
};
```

评论区精华

> gemini-code-assist[bot]（高优先级）：isAMD 检查在多处重复，建议使用 ['mi300x', 'mi325x', 'mi355x'].includes(hw) 替代多个 II。另外，生成器中为 AMD 禁用 FP8 但文档

> 说明 FP8 受支持且推荐，应使生成器与文档一致。
>
> gemini-code-assist[bot]（中优先级）：投机解码条件和命令生成中的 isAMD 同样建议用 .includes() 保持一致性。

作者采纳了 .includes() 重构建议，但未开启 AMD FP8 支持，该不一致仍有待后续解决。

风险与影响

风险：极低。所有变更仅限于文档 UI 组件和配置，不涉及任何后端逻辑或模型推理路径。MI325X 行为与之前合并选项完全一致。

影响：用户将获得更精确的硬件选择，导航顺序更符合预期（新版模型优先）。团队维护成本极低。

关联脉络

该 PR 与 #23532（添加 Hunyuan 3 Preview cookbook）同属文档站点的持续更新，均涉及 docs_new 下的卡片和导航配置。后续可关注 FP8 文档与生成器行为不一致的问题（#23540 中已记录但未解决），可能在其他 PR 中对齐。