PR #26511 完整报告

sgl-project/sglang

Update kimi k25 launch command in cookbook

合并时间：2026-05-28 07:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26511

执行摘要

本 PR 更新了 Kimi K25 部署指南中的启动命令，将 speculative draft 模型路径指向 lightseekorg/kimi-k2.5-eagle3-mla，并为 Blackwell B300 硬件新增了 tokenspeed_mla attention 后端配置。仅涉及一个 JSX 文件，改动量小，属于快速文档维护。

功能与动机

为了让用户能够正确部署 Kimi K25 模型，需要更新 cookbook 中的启动命令以匹配实际使用的模型权重和硬件支持。主要解决了以下问题：

- speculative draft 模型路径更新（新增 -mla 后缀）
- 为 Blackwell B300 硬件提供正确的 attention 后端配置

实现拆解

1. 更新 speculative draft 模型路径：将 speculative-draft-model-path 参数从 lightseekorg/kimi-k2.5-eagle3 修改为 lightseekorg/kimi-k2.5-eagle3-mla。
2. 新增 Blackwell B300 条件分支：当用户选择 B300 硬件时，在命令中添加 --attention-backend tokenspeed_mla 参数。
3. 改动仅影响命令字符串拼接逻辑，不影响组件结构、样式或交互行为。

docs_new/src/snippets/autoregressive/kimi-k25-deployment.jsx

唯一变更文件，更新了 speculative draft 模型路径并新增 Blackwell B300 attention 后端配置。

关键源码片段

docs_new/src/snippets/autoregressive/kimi-k25-deployment.jsx

唯一变更文件，更新了 speculative draft 模型路径并新增 Blackwell B300 attention 后端配置。

```
    // 在 KimiK25Deployment 组件的命令构建函数中：

    // Speculative decoding (EAGLE3)
    if (speculative === 'enabled') {
      cmd += ' \\
  --speculative-algorithm EAGLE3' +
        ' \\
```

```
--speculative-num-steps 3' +
      ' \\
--speculative-eagle-topk 1' +
      ' \\
--speculative-num-draft-tokens 4' +
      // 注意：模型路径已从 lightseekorg/kimi-k2.5-eagle3 更新为带 -mla 后缀的新模型
      ' \\
--speculative-draft-model-path lightseekorg/kimi-k2.5-eagle3-mla';
  }

  // 新增：Blackwell (B300) 专用 tokenspeed MLA attention 后端
  // 当用户选择 B300 硬件时，自动追加该参数以启用优化的 attention 实现
  if (hardware === 'b300') {
    cmd += ' \\
--attention-backend tokenspeed_mla';
  }
```

评论区精华

无讨论。

风险与影响

风险：极低。仅修改文档示例命令，但应确保新模型路径和 attention 后端在实际环境中可用，避免用户部署失败。

影响：仅影响访问 cookbook 页面的 Kimi K25 用户，使其获得更准确的部署命令。对于 Blackwell B300 用户，自动追加正确的 attention 后端参数，简化部署配置。

关联脉络

无关联的 Issue 或历史 PR。