

PR #34886 完整报告

sgl-project/sglang

[Docs] Update Kimi-K3 installation options

合并时间: 2026-08-15 07:19

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34886>

执行摘要

PR 34886 是纯文档变更, 仅修改 1 个文件 (+14/-8), 将 Kimi-K3 cookbook 的安装说明从预发布专有镜像 (kimi-k3、kimi-k3-cu12、ROCm daily) 和 kimi-k3 分支 Dockerfile 回退, 切换为标准的 Python (pip / uv) 与 lmsysorg/sglang:latest Docker 安装路径。目的在于让 K3 安装文档与 SGLang 当前发布体系对齐, 无任何运行时影响。

功能与动机

Kimi-K3 的 cookbook 此前仍引导读者使用预发布专有镜像和 kimi-k3 分支的 Dockerfile。PR body 明确指出:

Now that Kimi-K3 is supported by the current SGLang release paths, the installation section should use the standard Python and Docker instructions.

即 K3 已进入标准发布路径, 安装文档应回归官方统一说明, 避免读者继续依赖预发布镜像或源码构建路径。

实现拆解

1. 变更入口: 唯一改动文件为 docs/cookbook/autoregressive/Moonshotai/Kimi-K3.mdx, 改动集中在 Install SGLang 的 Accordion 折叠面板内, 不改变页面整体布局和命令面板结构。
2. 新增 Python 安装路径: 在原有 Docker 标签页前新增 Python (pip / uv) 标签页, 给出 pip install --upgrade pip、pip install uv、uv pip install sglang 三步命令, 并提示用户使用下方命令面板生成的 Python 启动命令, 使安装说明与部署命令面板的 Python / Docker 开关一一对应。
3. 统一 Docker 镜像标签: 删除 kimi-k3 (CUDA13)、kimi-k3-cu12 (CUDA12) 和 ROCm daily 三个预发布标签, 替换为 lmsysorg/sglang:latest, 并保留指向完整安装指南 Method 3 的 Docker 启动方式说明。
4. 清理预发布残留: 移除“直到这些标签随公开 K3 发布”的过时声明, 以及指向 kimi-k3 分支两个 Dockerfile 的回退指引, 防止用户误用源码构建路径。
5. 配套验证: 本 PR 为纯文档变更, 未增删测试; 作者使用 node docs/scripts/check_cookbook_configs.mjs、mint validate、mint broken-links 三组命令校验文档配置、格式和链接完整性。

<docs/cookbook/autoregressive/Moonshotai/Kimi-K3.mdx>

本 PR 唯一变更文件，集中改写了 Kimi-K3 安装说明，是文档更新的全部内容载体。

```
<Accordion title="Install SGLang">
  <!-- 安装说明与命令面板的 Python / Docker 开关保持一一对应 -->
  <Tabs>
    <Tab title="Python (pip / uv)">
      <!-- 使用 uv 统一管理 Python 依赖，再运行下方命令面板生成的 Python 启动命令 -->
      ```bash
pip install --upgrade pip
pip install uv
uv pip install sglang
```

      Then run the Python output of the command panel below in that environment.
    </Tab>

    <Tab title="Docker">
      <!-- 直接使用当前发布镜像，取代预发布的 kimi-k3 / kimi-k3-cu12 等标签 -->
      ```bash
docker pull lmsysorg/sglang:latest
```

      For how to launch the image, see
      [Install → Method 3: Using Docker](../../docs/get-started/install#method-3-using-docker).
    </Tab>
  </Tabs>
</Accordion>
```

评论区精华

本 PR 没有任何 review 评论或讨论线程，作者自审并直接合入，属于低争议的文档收尾变更。

风险与影响

- 文档一致性风险：文档将 Docker 标签固定为 lmsysorg/sglang:latest，若未来某版本镜像不包含 Kimi-K3 支持，文档会过时；broken-links 只能校验链接有效性，不能校验内容语义。
- 用户迁移风险：原 CUDA12、ROCm daily 预发布镜像用户切换 latest 后，需自行确认对应硬件平台镜像仍可用。
- CI 观察项：PR 的 Extra 测试标记失败（Run #31849432967）且未说明原因；由于本 PR 不触碰运行时代码，该失败大概率与本次文档变更无关，但合入时可留意。
- 影响面：仅影响文档读者与维护者，无运行时影响。

关联脉络

- 与 #34741 的关联：该 PR 修复 AMD gfx950 上 Triton 3.7 extend-attention 的 spill 问题并恢复 Kimi-K3 吞吐，说明 K3 的适配与性能工作与本文档收尾同期进行，可能预示 K3 即将进入正式发布渠道。
- 与 #34860、#31554 的关联：均为 cookbook 文档维护工作（新增部署页、填充基准数据），反映 SGLang 持续对模型 cookbook 做标准化整理。

- 后续关注：K3 正式发布后可能还有基准表、硬件矩阵等配套文档 PR，本文档收敛是这一流程的组成部分。