# PR #25813 完整报告

sgl-project/sglang

docs(cookbook): port popular model usage guides into cookbook pages

合并时间：2026-06-02 08:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25813

## 执行摘要

- 一句话：将 11 个旧文档页面的模型用法移植到 cookbook
- 推荐动作：值得精读，展示了大规模文档迁移的完整流程和审核实践，尤其适合需要合并整理知识库的场景。

## 功能与动机

关联 issue #141 要求合并 sglang 文档到 cookbook。PR body 说明需要将 docs/basic_usage/ 下 11 个模型用法页面内容移植到 cookbook，移除对旧文档的所有引用，使用户只需访问 cookbook 即可获得完整的模型部署信息。

## 实现拆解

1. 对每个旧文档页面（如 deepseek_v3.mdx, deepseek_v32.mdx, glm45.mdx 等）逐一将硬件配置、启动命令、配置提示等核心内容移植到对应的 cookbook 页面（如 DeepSeek-V3.mdx, DeepSeek-V3_2.mdx, GLM-4.5.mdx 等）。
2. 修正交互式命令生成器（JSX 组件）中的模型路径错误，例如 llama4-maverick-deployment.jsx 从 Scout 改为 Maverick。
3. 更新重定向脚本 gen_redirects.py 和导航配置 docs.json，使旧链接指向新的 cookbook 位置。
4. 在审核过程中修复大量内容错误，包括移除不存在的 CLI 标志、修正模型 ID、统一端口号、修复死链接等。
5. 最终删除所有旧文档页面，完成迁移。

关键文件：

- docs_new/src/snippets/autoregressive/llama4-maverick-deployment.jsx（模块 代码片段；类别 source；类型 core-logic）：修正了模型路径从 Scout 到 Maverick，并增加了硬件和量化选项，是交互式命令生成器的核心修正。
- docs_new/scripts/gen_redirects.py（模块 部署脚本；类别 source；类型 core-logic）：更新了从旧 basic_usage 页面到 cookbook 页面的 HTTP 重定向映射，确保旧链接有效。
- docs_new/cookbook/autoregressive/GLM/GLM-4.6V.mdx（模块 文档页面；类别 other；类型 core-logic；符号 image_to_base64）：接收了来自 glmv.mdx 的移植内容，包括多模态参数、启动命令，并修复了死链接和格式问题。

- docs_new/docs/basic_usage/deepseek_v32.mdx（模块 旧文档；类别 other；类型 deletion）：被删除的旧文档之一，内容已迁移到 cookbook。
- docs_new/cookbook/autoregressive/Llama/Llama4.mdx（模块 文档页面；类别 other；类型 core-logic）：接收了来自 llama4.mdx 的配置提示和 EAGLE 推测解码内容。

关键符号：image_to_base64

## 关键源码片段

### docs_new/src/snippets/autoregressive/llama4-maverick-deployment.jsx

修正了模型路径从 Scout 到 Maverick，并增加了硬件和量化选项，是交互式命令生成器的核心修正。

```
export const Llama4MaverickDeployment = () => {
  // 硬件选项增加 B200、H200，默认仍为 MI300X
  const options = {
    hardware: {
      items: [
        { id: 'b200', label: 'B200', default: false },
        { id: 'h200', label: 'H200', default: false },
        { id: 'mi300x', label: 'MI300x', default: true },
        { id: 'mi325x', label: 'MI325x', default: false },
        { id: 'mi355x', label: 'MI355x', default: false }
      ]
    },
    quantization: { /* BF16 默认，FP8 可选 */ },
    toolcall: { /* 工具调用解析器开关 */ },
    speculative: { /* EAGLE3 推测解码开关 */ },
    host: { /* 默认 0.0.0.0 */ },
    port: { /* 默认 8000 */ }
  };

  // 生成启动命令：关键修复 --model-path 从 Scout 改为 Maverick
  const generateCommand = (values) => {
    let cmd = 'python -m sglang.launch_server \\
';
    cmd += `  --model-path meta-llama/Llama-4-Maverick-17B-128E-Instruct`;
    // 根据硬件设置 TP 大小
    if (hardware === 'h200' || hardware === 'b200') cmd += ` \
  --tp 8`;
    // EAGLE3 推测解码时使用对应的 Maverick 草稿模型路径
    if (speculative === 'enabled') {
      cmd += `  --speculative-draft-model-path lmsys/sglang-EAGLE3-Llama-4-Maverick-17B-128E-Instruct-v1`;
    }
    // ... 其余固定参数
    return cmd;
  };
};
```

### docs_new/scripts/gen_redirects.py

更新了从旧 basic_usage 页面到 cookbook 页面的 HTTP 重定向映射，确保旧链接有效。

```
# 重定向字典：旧路径 -> 新 cookbook 路径
EXPLICIT = {
    # ... 其他映射
    # 将所有 basic_usage 页面指向对应的 cookbook 页面（而非旧位置）
    "/basic_usage/kimi_k2_5": "/cookbook/autoregressive/Moonshotai/Kimi-K2.5",
    "/basic_usage/deepseek_ocr": "/cookbook/autoregressive/DeepSeek/DeepSeek-OCR",
    "/basic_usage/deepseek_v3": "/cookbook/autoregressive/DeepSeek/DeepSeek-V3",
    "/basic_usage/deepseek_v32": "/cookbook/autoregressive/DeepSeek/DeepSeek-V3_2",
    "/basic_usage/glm45": "/cookbook/autoregressive/GLM/GLM-4.5",
    "/basic_usage/glmv": "/cookbook/autoregressive/GLM/GLM-4.6V",
    "/basic_usage/gpt_oss": "/cookbook/autoregressive/OpenAI/GPT-OSS",
    "/basic_usage/llama4": "/cookbook/autoregressive/Llama/Llama4",
    "/basic_usage/minimax_m2": "/cookbook/autoregressive/MiniMax/MiniMax-M2",
    "/basic_usage/popular_model_usage": "/cookbook/autoregressive/intro",
    "/basic_usage/qwen3": "/cookbook/autoregressive/Qwen/Qwen3",
    "/basic_usage/qwen3_5": "/cookbook/autoregressive/Qwen/Qwen3.5",
    "/basic_usage/qwen3_vl": "/cookbook/autoregressive/Qwen/Qwen3-VL",
    # ... 其他映射
}
```

## 评论区精华

主要讨论围绕内容的准确性和完整性。审核者 zijiexia 提出 50 多项问题，分为 P0-P3 优先级：

- P0 问题如引入了不存在的 --mm-max-concurrent-calls 标志、DeepSeek-R1 的思维预算被错误放在 V3 页面、OCR 示例使用错误模型 ID、链接失效等。作者逐一修正。
- 多次回合要求补充缺失的配置提示（如 MTP 推测解码部分）和调整标题层级。
- 最终在作者完成所有修复后，审核者批准合并。
- 移除不存在的 CLI 标志 (correctness): 作者确认并从三个文件中删除了这些标志，替换为通用说明。
- 思维预算位置错误 (correctness): 作者将整个 §4.2.3 块从 V3.mdx 移到 R1.mdx，并调整编号。
- 重定向和导航配置不完整 (design): 作者在 gen_redirects.py 中添加了所有 11 个页面的重定向映射，并修正了 doc.json 中的导航和 HTML 重定向。

## 风险与影响

- 风险：PR 是纯文档变更，无代码影响，风险较低。但内容迁移过程中可能引入错误或遗漏，通过详细审核已修复。主要风险是重定向配置错误或旧链接失效，已通过验证确保正确。
- 影响：用户影响：之前位于 basic_usage 的模型用法指南现在全部在 cookbook 中，用户需要导航到新的位置。系统影响：无。团队影响：后续文档维护集中在 cookbook，旧页面不再需要更新。

- 风险标记：内容迁移错误，重定向配置遗漏

## 关联脉络

- 暂无明显关联 PR