

PR #34658 完整报告

sgl-project/sglang

[do not merge] add new cookbooks

合并时间: 2026-08-13 21:35

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34658>

执行摘要

- 一句话: 新增 Dots3-Note 部署 cookbook, 将 #33829 配方文档化
- 推荐动作: 作为代码 PR 无需深入阅读, 但值得文档维护者参考其“配置驱动部署面板”的结构: jsx 单一 config 字面量与 `_deployment.jsx` 的组合, 让模型 cookbook 的部署命令可维护、可参数化。若后续跟进, 建议在 checkpoint 发布后替换 `modelName.default`、补上 HF 链接、移除 `verified: false` 并补充真实基准数据。

功能与动机

PR body 明确说明是“Follow-up docs for #33829: converts the dots.note.omni serving recipe into the standard cookbook MDX format.”, 即把 #33829 中验证过的 dots3.note 部署配方 (单节点 8xH200、DP8xTP8xEP8、NEXTN 投机、hybrid MLA/SWA) 沉淀为可复现的标准 cookbook 页面, 降低用户部署门槛; 同时因为 checkpoint 尚未公开, 页面以 #33829 分支作为验证基线。

实现拆解

实现拆解如下:

1. 新增部署配置驱动文件 `docs/src/snippets/configs/rednote/dots3-note.jsx` (+165 行): 单一 `export const config` 字面量, 无 spread/call/IIFE, 注释明确说明这是为了适配 Mintlify hydration 时重新求值的约束。配置定义 `modelName`、`showPlaygroundLink: false` (dots tool-call parser 已内置于 cells)、`supportedHardware: ["h200"]` (暂不支持 Blackwell)、`matchDims` (唯一部署选择为 BF16/FP8 精度)、`placeholders` (HOST_IP/PORT/HF_TOKEN/CURL_HOST/CURL_PORT 五种占位符, 按 target 区分替换进 command 或 curl)、curl 多模态示例 (video_url + text)、Docker 镜像 `lmsysorg/sglang:dev`, 以及两个部署单元 (bf16 与 fp8), 每个单元含 env 与完整 serving flags。该文件被页面中的 `<Deployment config={config} />` 消费, 用于渲染部署命令面板。
2. 新增 cookbook 页面 `docs/cookbook/autoregressive/RedNote/Dots3-Note.mdx` (+213 行): 包含 front matter (title/description/tag: NEW)、Install SGLang 手风琴 (pip/uv 与 Docker 两个安装路径, Docker 推荐 `dev-dots3-note` 镜像)、部署单元说明 (精度选择、NEXTN 投机、MTP 全共享、FA3 吸收式 SWA-MLA)、模型简介 (MoE ViT、Whisper-derived 音频编码器、native video 展平管线、hybrid attention、DSA、MTP)、配置技巧 (混合 KV 池 `--swa-full-tokens-ratio 0.03`、MoE runner、FA3、DSA 关闭方法、

CUDA graph、context length、language-only 模式），以及高级用法中的 native video 输入示例与 `extra_body` 参数表（seq/audio_cap/audio_sr/k_mode）。所有部署单元均标记为未验证。

3. 注册站点导航 `docs/docs.json` (+6 行)：在 autoregressive cookbook 目录中新增 RedNote group，指向 `cookbook/autoregressive/RedNote/Dots3-Note` 页面，插入位置位于 Meituan 与 Google 分组之间。
4. 提交演进：4 个 commit 依次完成“新增 cookbook”、“合并 jianfei-wangg 分支修正 Dots3-Note 命名与 serving 建议”、“推荐 dev-dots3-note Docker 镜像”、“修复 EPD anchor 链接”，反映内容在评审前经过一轮协作修正。
5. 测试与配套：本次为纯文档变更，没有测试文件；CI 记录显示 extra 测试失败（Run #31705003890），但仓库内没有对应评论可以归因原因。

关键文件：

- `docs/src/snippets/configs/rednote/dots3-note.jsx`（模块 文档片段；类别 source；类型 core-logic；符号 config）：cookbook 页面的核心驱动文件，定义 Deploy 面板的全部部署单元、占位符与 curl 示例；被 `_deployment.jsx` 消费，是本次文档功能的关键实现单元。
- `docs/cookbook/autoregressive/RedNote/Dots3-Note.mdx`（模块 模型文档；类别 docs；类型 documentation）：cookbook 正文页面，承载安装引导、模型介绍、配置技巧与 native video 用法，是用户实际阅读的文档主体。
- `docs/docs.json`（模块 站点导航；类别 config；类型 configuration）：文档站点导航配置，将新页面挂载到 autoregressive cookbook 的 RedNote 分组，是页面可被访问的必需配套。

关键符号：未识别

关键源码片段

`docs/cookbook/autoregressive/RedNote/Dots3-Note.mdx`

cookbook 正文页面，承载安装引导、模型介绍、配置技巧与 native video 用法，是用户实际阅读的文档主体。

```
---
title: Dots3-Note
description: "Deploy RedNote dots3.note with SGLang — a native multimodal omni model (MoE ViT + Whisper-derived audio encoder + native video flattening) on the dots3 hybrid MLA/SWA language model, with DSA and full-sharing MTP speculative decoding."
tag: NEW
---
```

```
<!-- 部署小节：先引导安装含 #33829 的 SGLang 构建，再渲染配置驱动的 Deploy 面板 -->
```

```
## Deployment
```

```
<Accordion title="Install SGLang">
```

```
dots3.note support is in [SGLang PR #33829](https://github.com/sgl-project/sglang/pull/33829).
```

Until that PR is included in a tagged SGLang release, install from a build that contains the PR.

<Tabs>

<Tab title="Python (pip / uv)">

```
```bash Command
pip install -U uv
uv venv --python 3.12 && source .venv/bin/activate
git clone https://github.com/sgl-project/sglang.git
cd sglang
git fetch origin pull/33829/head && git checkout FETCH_HEAD
uv pip install -e python
```

```
```bash Command docker pull lmsysorg/sglang:dev-dots3-note <!-- dev-dots3-note 镜像
已打包 #33829 的 dots3.note 支持，省去自行构建分支 --> </Tab> </Tabs> </Accordion>
<!-- 部署面板由下方 config 驱动，支持 BF16/FP8 两种精度单元；当前全部标记为未验证
--> import { Deployment } from "/src/snippets/_deployment.jsx"; import { config } from "
/src/snippets/configs/rednote/dots3-note.jsx"; <Deployment config={config} /> <Note>
Every cell in the Deploy panel above is currently unverified: the recipe runs, but no
serving round on public weights has landed (the checkpoint is not yet released). Treat
the cells as starting points and re-measure throughput and accuracy on your workload.
</Note>
```

评论区精华

该 PR 没有任何 review 评论 (review_comments_count = 0)，仅有一条 JustinTong0323 的空 body APPROVED。标题带有 “[do not merge]” 前缀但最终被维护者合入，说明该前缀在此处是流程占位而非真实阻止合并的意图。由于缺少讨论，无法呈现设计交锋；从提交历史看，jianfei-wangg 的合入提交（修正命名与 serving 建议）和后续“推荐 dev-dots3-note 镜像”“修复 epd anchor”提交属于提交层面的自审修正。

- 暂无高价值评论线程

风险与影响

- 风险：风险主要集中于文档真实性与可执行性，具体如下：
 1. 占位模型名不可用：docs/src/snippets/configs/rednote/dots3-note.jsx 中 modelNames.default 是占位符 <dots-note-checkpoint>，cells 的 --model-path {{MODEL_NAME}} 与 --speculative-draft-model-path 都会引用它；在 checkpoint 公开发布前，用户直接复制命令必然失败。
 2. 部署单元未验证：两个 cells 的 verified: false，且页面明确提示 recipe “runs but no serving round on public weights has landed”，其中的吞吐与参数建议（如 --mem-fraction-static 0.87、--cuda-graph-max-bs-decode 32）未经公开权重基准验证，可能随权重发布后调整。

3. 强依赖 #33829: 安装说明要求从 pull/33829/head 构建或拉取 lmsysorg/sglang:dev-dots3-note 镜像; 若对应镜像未构建或 #33829 长期未合入 tag, Docker 路径会失效。
4. Mintlify 配置约束: jsx 文件头注释强制“单一 export const config 字面量”, 后续维护者若加入 spread/call/IIFE, 会在 hydration 阶段破坏 Deploy 面板渲染。
5. 导航路径一致性: docs/docs.json 新增的 cookbook/autoregressive/RedNote/Dots3-Note 路径必须与 MDX 文件实际路径保持一致, 否则文档站点会出现 404; tag: NEW 还依赖站点对 new 标签的展示逻辑。
6. CI extra 失败未知: PR 状态区显示 extra 测试 Run #31705003890 失败, 但没有注释或讨论归因, 无法判断是文档构建问题还是无关的 CI 波动。
 - 影响: 对用户: 提供 dots3.note 唯一的官方部署手册入口, 在 checkpoint 发布前是“预告文档”, 发布后可直接照 cookbook 一键生成部署命令; 对普通 SGLang 用户, 导航多出一个 RedNote 分组, 不影响既有页面。对系统: 零运行时影响, 纯文档站点 (Mintlify) 资产变化。对团队: 新增一条 RedNote 模型文档线, 后续 RedNote 系模型可复用同一 docs/src/snippets/configs/rednote/ 目录的配置驱动模板; 同时要求维护者跟踪 #33829 合入与 checkpoint 发布后替换占位符。
 - 风险标记: 部署单元均未验证, 依赖未发布 checkpoint, 绑定 PR #33829, Mintlify 配置约束

关联脉络

- PR #33829 dots3.note model support (标题依据 PR body 推断): 本 PR 的 body 明确声明是 #33829 的 follow-up docs; 页面安装说明、资源链接与验证基线都指向该 PR, checkpoint 发布前所有部署单元以它为准。