

PR #34247 完整报告

sgl-project/sglang

[Docs] Standardize diffusion cookbook model pages

合并时间: 2026-08-21 10:25

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34247>

执行摘要

本 PR 把 diffusion cookbook 模型页从各自为政的部署矩阵，收敛为「能力标签 + Quick start + 能力 / 选型 / 边界说明」的统一契约，并在共享部署命令组件 `_deployment.jsx` 中新增 `commandBuilder` 渲染路径，让 MiniMax-H3 等模型页按生命周期 (base/serve/request) 组织配置。配套新增 authoring 模板与 `check_cookbook_configs.mjs` 结构校验，使后续模型页自动遵循同一合同。整体是文档站基础设施的一次向前兼容扩展：旧 config 走原渲染器，新 config 走响应式 builder。

功能与动机

Diffusion cookbook 页面开头风格不一致，且注意力、量化、缓存、编码器调度等正交运行时特性混在部署矩阵里，让读者难以判断模型能力边界。PR body 明确: "This keeps base recipes small while making feature quality contracts and verified scope explicit." 目标是把基础配方保持精简，同时显式化 feature 质量契约与已验证范围，让读者更快识别模型优势、选择合适的 checkpoint 与请求模式，并单独应用可选性能特性。

实现拆解

1. 定义作者契约与模板：新增 `.claude/skills/cookbook-add-model/references/diffusion-authoring.md`、`diffusion-page.mdx.tmpl` 与 `diffusion-config.jsx.tmpl`，规定五节页面结构与 `base/serve/request` 三类 overlay 维度，让新增模型页从模板开始就符合同一合同。
2. 扩展共享引擎 `docs/src/snippets/_deployment.jsx`：新增 `commandBuilder` 数据契约，config 可声明 `defaultSelection`、`resource` (`limits/verifiedRecipes/autoTopology/validateTopology`) 与 `resolveDeployment`；同时引入软禁用档位 `soft/softReason`，区分「验证覆盖不足」与「能力边界」，并让触屏用户点击被禁选项时看到闪烁的 `disableReason`。
3. 重构 MiniMax-H3 配置：删除 `profile matchDims`，拓扑交给 `commandBuilder.resource`；按生命周期重组 overlay 维度并补齐 `docsHref/learnMore/quality/soft` 字段；修正 encoder DP 跨节点限制，把 BCG 等 `coverage-only` 门控移入 `soft` 档。
4. 配套样式与校验：`docs/custom.css` 新增 `.sgd-command-builder` 玻璃质感系统与 `container query`；`docs/scripts/check_cookbook_configs.mjs` 增加 builder 结构校验、数值维度 `bounds` 检查，并遍历 `cookbook/diffusion` 下的所有页面。
5. 页面标准化与分支整合：统一 README、MiniMax-H3、Cosmos3、FLUX、SANA-WM、Krea-2 等页面开头；与 `main` 分支两次合并，解决 `lint.yml` 与 `MiniMax-H3.mdx` 冲突。

`docs/src/snippets/_deployment.jsx`

共享部署命令生成引擎的核心文件，新增 commandBuilder 渲染路径与 soft 软禁用机制，是所有 cookbook 部署矩阵的公共底座。

```
// 部署命令生成器的共享引擎组件，被所有 cookbook 模型页复用。它只读取传入的
// config，不包含任何模型专属逻辑；字段语义与解析规则统一收口在
// .claude/skills/cookbook-add-model/references/authoring-reference.md 中。
//
// 两种选择维度：
// matchDims 参与 cell 查找，选中值决定命中哪条已验证命令；
// overlayDims 不参与查找，选中项把 flags/env/hints 叠加到命中的 cell 上，
// 因此正交旋钮（如编码器并行）不会让部署矩阵成倍膨胀。
// 新增的 commandBuilder 字段把 config 切换到响应式 diffusion 构建器：它复用
// 本引擎的 overlay 组合与命令渲染，但由 config 自己提供 defaultSelection、
// resource（verifiedRecipes + 拓扑自动推导 / 校验）以及 resolveDeployment，
// 后者返回 cell 外加 builder 元数据（topologySummary、errors、warnings、
// verification、resolvedSettings）。UI 独有的 scope/expand 与本地
// head-address/rank 状态不会进入 URL hash。

export const Deployment = ({ config, benchmarks }) => {
  // 有 commandBuilder 时启用新的响应式渲染路径，旧 config 始终保持原渲染器
  const commandBuilder = config.commandBuilder || null;

  // DIMENSIONS 按优先级排序：高下标维度适配低下标维度的已选值，反之不行。
  // 该顺序驱动选项的置灰 / 吸附逻辑。
  const DIMENSIONS = ["hw", ...matchDimSpecs.map((d) => d.id)];

  // `soft` 标记“可能可行但处于已验证矩阵之外”的选项：它保持可选，验证状态
  // 由徽章单独表达；`disabled` 只留给真正无法工作的组合。触屏设备无法悬浮
  // 查看 title 提示，所以被阻止的选项在被点击时会在行下方闪烁显示
  // disableReason。
  const optionSoft = (opt, sel) =>
    typeof opt.soft === "function" ? opt.soft(sel) : !!opt.soft;

  // 其余辅助逻辑（normalizeBuilderSelection、flashBlockedNote、builderMeta
  // 等）与既有 cells/benchmark 渲染共用同一个 selection 对象。
};
```

评论区精华

PR 没有正式 review 评论，但 commit 历史揭示了来自 review 反馈的迭代（commit cce957f 提到 "Three fixes from review"）：

"Recommended for B300" was misinformation: the 8-GPU shape is what the verification round happened to run, not sizing advice (B300 at 288GB does not need more GPUs than H200 at 141GB).

Classify every disabled predicate in the H3 config by what it encodes. Capability boundaries keep hard blocks; verification coverage becomes the soft tier, so the option stays selectable and reports itself as unverified.

The cookbook blocked encoder DP on multi-node deployments, but the runtime gate only requires `tp_size=1` and `dp_size=1` — the DP group is the world group, which spans nodes.

另外 CI bot 显示 Mintlify 预览就绪，PR Test (Extra) 曾失败，作者以 `/tag-and-rerun-ci` 重跑后合并。

风险与影响

风险：

- `_deployment.jsx` 是文档站部署矩阵的公共引擎，+803 行修改存在回归风险，旧 config 虽走原渲染器，但 DIMENSIONS、optionSoft 等公共逻辑变化仍可能影响既有模型页。
- `docs/custom.css` +666 行全局样式，选择器虽有范围限定，但 `.sgd-model-tags` 等旧规则被改动，可能影响其他页面布局。
- `check_cookbook_configs.mjs` 的强制校验可能将「合法但未预期」的配置判为失败，脚本与模板需保持同步。
- 本次改动没有直接测试文件，依赖 pre-commit 与 CI 脚本兜底。
- 39 个 commit、两次 merge main 解决冲突，存在内容遗漏的潜在风险。

影响：读者侧获得更清晰的模型选型体验；维护者侧新增文档作者有模板与 CI 校验托底；系统侧文档站增加一个长期维护的 `commandBuilder` 渲染路径，对运行时无影响。

关联脉络

本 PR 与近期文档站 PR 形成连续演进：PR#35753 与 PR#35663 围绕 Qwen3.8-27B DFLASH2 cookbook，复用 `_deployment.jsx` 与 `check_cookbook_configs.mjs` 的既有能力；本 PR 则把该组件从 autoregressive 场景扩展到 diffusion 场景，并补上了「模板 + 校验 + skill 文档」的完整作者流程。整体方向是把 cookbook 部署矩阵从模型作者手工维护的二维表格，演进为数据契约驱动的可校验交互组件，未来新增模型页的成本会显著下降。