

PR #27845 完整报告

sgl-project/sglang

docs: add cookbook-migrate-model skill from the Qwen3.5 pilot

合并时间: 2026-06-13 09:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27845>

执行摘要

- 一句话: 添加 cookbook 迁移技能, 规范遗留模板迁移
- 推荐动作: 建议文档维护者快速接管本 PR, 同时注意:
 - 仔细审阅 SKILL.md 中所有硬规则和 dimension-mapping.md 中的映射表, 确保覆盖所有遗留模式。
 - 关注策略命名规则和精度标志处理的设计权衡, 这些是 cookbook 数据模型的关键决策。
 - 后续迁移工作应严格遵循本技能定义的工作流。

功能与动机

PR #27848 迁移了第一个遗留模板 cookbook (Qwen3.5), 本 PR 将这次经验提炼为可复用的 `cookbook-migrate-model` 技能, 使其余遗留页面 (GLM、Kimi、MiniMax、Qwen 等家族) 的迁移可以由任何贡献者或代理一致地执行, 无需每次重新推演约定。

实现拆解

1. 创建迁移技能核心: 新增 `.claude/skills/cookbook-migrate-model/SKILL.md`, 定义完整迁移工作流、硬规则 (不现代化、不发明版本、精度标志处理等) 和策略关联规则。
2. 添加维度映射参考: 新增 `references/dimension-mapping.md`, 提供遗留控制到新维度的映射表、策略集定义、命令规范化规则, 以及 Qwen3.5 决策日志。
3. 更新审查技能: 修改 `.claude/skills/cookbook-review-pr/SKILL.md`, 补充 playground 轴 opt-out 原则、策略计数检查、精度标志审查项, 以及准确度基准一致性检查。
4. 更新添加模型技能模板: 修改 `config.jsx.tmpl` 和 `page.mdx.tmpl`, 应用 playground 轴 opt-out、策略三合一 (low-latency/balanced/high-throughput)、高级用法可折叠等约定。同步更新 `authoring-reference.md` 等参考文档。
5. 细化策略与精度规则: 通过多次提交逐步完善策略命名 (禁止模型特定 id)、精度退化标志默认不嵌入 cells、MTP 到策略的确定性映射等关键决策。
6. 更新迁移轮次清单: 将迁移范围缩小到 8 个模型, 并添加 Gemma4 和 Nemotron3-Ultra 的家族表行。

关键文件:

- `.claude/skills/cookbook-migrate-model/SKILL.md` (模块 迁移技能; 类别 docs; 类型 documentation): 核心技能定义, 包含完整迁移工作流程、硬规则和策略映射规则。

- `.claude/skills/cookbook-migrate-model/references/dimension-mapping.md` (模块 维度映射参考; 类别 docs; 类型 documentation) : 提供遗留控制到新维度的详细映射表、策略集和命令规范化规则, 是迁移工作的核心参考。
- `.claude/skills/cookbook-review-pr/SKILL.md` (模块 审查技能; 类别 docs; 类型 documentation) : 更新了审查检查清单, 补充 playground 轴 opt-out、策略命名、精度标志等审查点。
- `.claude/skills/cookbook-add-model/templates/config.jsx.tmpl` (模块 配置模板; 类别 other; 类型 data-contract) : 配置模板添加策略字段规则和精度标志注释, 应用 opt-out 哲学。
- `.claude/skills/cookbook-add-model/templates/page.mdx.tmpl` (模块 页面模板; 类别 other; 类型 data-contract) : 页面模板调整高级用法部分为可折叠 Accordion 格式, 符合需求。
- `.claude/skills/cookbook-add-model/references/authoring-reference.md` (模块 创作参考; 类别 docs; 类型 documentation) : 更新策略字段说明和精度标志处理规则, 与迁移技能保持一致。
- `.claude/skills/cookbook-add-model/references/mintlify-authoring.md` (模块 Mintlify 参考; 类别 docs; 类型 documentation) : 同步更新 MDX 编写规则。
- `.claude/skills/cookbook-add-model/SKILL.md` (模块 添加模型技能; 类别 docs; 类型 documentation) : 添加指向迁移技能的交叉引用。
- `.claude/skills/cookbook-add-model/references/engine-axis.md` (模块 引擎轴参考; 类别 docs; 类型 documentation) : 同步更新引擎轴参考文件。

关键符号: 未识别

评论区精华

本 PR 无直接 review 评论, 但通过提交历史可见以下关键设计决策:

- 策略命名规则: 策略 id 必须复用 low-latency/balanced/high-throughput 词汇, 禁止模型特定 id (如 mtp/no-mtp) 。
- Playground 轴 opt-out: 一般轴默认包含在每一个 cookbook 页面, 仅当模型确实无法使用时删除, 而非 opt-in。
- 精度退化标志处理: 低于检查点精度的运行时量化 (如 W4A4、lossy KV cache dtype) 默认不嵌入 cells, 仅作为 Playground 选项或提示, 除非在无条件命令中被测量。
- MTP 策略映射: MTP/speculative decoding 开启映射到 **low-latency**, 关闭映射到 **high-throughput**, 作为确定性默认, 反向映射需要维护者确认。这些决策通过连续提交逐步完善, 确保了迁移工具的一致性和可维护性。
- 策略命名禁止使用模型特定 id (design): 采纳该规则, 更新 SKILL.md 和 config.jsx.tmpl。
- Playground 轴由 opt-in 改为 opt-out (design): 采用 opt-out 模式, 更新 review 技能和 config 模板。
- 精度退化标志嵌入细胞的规则 (correctness): 精度退化标志 (如 W4A4、lossy KV cache dtype) 默认不嵌入 cells, 仅作为 Playground 选项或提示。

- MTP 到策略的默认映射 (design): 确立确定性默认映射, 减少迁移歧义。

风险与影响

- 风险: 本 PR 仅涉及文档和技能文件变更, 无代码风险。但技能规则的完整性和一致性直接影响未来 18 个遗留 cookbook 页面的迁移质量。任何遗漏或矛盾可能导致迁移结果不一致, 增加返工成本。特别需要关注指令覆盖完整性 (如所有 legacy 控件映射到新维度的规则覆盖)、策略命名强约束、精度标志处理规则的确定性。
- 影响: 影响范围: 对文档维护者和 AI 代理的 cookbook 迁移 workflow 产生直接影响。迁移技能将作为标准化工具, 降低人工迁移成本并保证输出一致性。当前遗留页面迁移轮次包含 10 个模型 (GLM、Kimi、MiniMax、Qwen、Gemma、Nemotron 等家族), 将统一采用本 PR 定义的规则。对终端用户无直接影响。

影响程度: 中等, 影响文档团队开发流程, 但不会影响运行时或 API。

- 风险标记: 文档变更, 影响迁移流程, 规则一致性依赖

关联脉络

- PR #27848 cookbook: migrate Qwen3.5 onto the config-driven template: 本 PR 基于 #27848 的迁移经验提炼技能, 是直接前驱。
- PR #28087 [Doc] Fix some inconsistencies in the Nemotron Cookbook: 相关 cookbook 文档修复, 与本 PR 同样维护 cookbook 体系。
- PR #25881 Fix Responses API request handling: 虽然不直接关联, 但同属于 API/ 文档方向, 体现了持续完善文档生态的趋势。