

PR #2381 完整报告

radixark/miles

docs: recipe pages for Kimi-K3, Nemotron-3-Ultra and Gemma-4

合并时间: 2026-08-12 07:04

原文链接: <http://prhub.com.cn/radixark/miles/pull/2381>

执行摘要

- 一句话: 补齐 Kimi-K3、Nemotron-3-Ultra、Gemma-4 配方文档
- 推荐动作: 值得精读, 尤其是 docs/models/nemotron/nemotron-3-ultra.md 对 `n_groups=8` 如何约束整个并行布局的分析, 以及 docs/models/kimi/kimi-k3.md 关于 LoRA 适配 MXFP4、CUDA IPC 权重同步与 `lora_base_cpu_backup` 的工程细节——这些是文档里最接近设计文档的内容。对团队的意义在于确立了 recipe 页必须只引用已存在资源、并与 scripts/ 保持一致内容的约定。

功能与动机

PR body 明确指出“Brings the model docs in line with what scripts/ actually ships”, 即文档与仓库实际提供的脚本脱节: Gemma-4 26B-A4B、Gemma-4 31B、JoyAI-LLM-Flash、Nemotron-3-Ultra-550B-A55B 早已有 launch 脚本且已在首页广告, 但在支持模型总表中缺失、且没有 recipe 页面, 只能链向裸脚本或 PR 链接; Kimi-K3 同样只有 #1825 的 PR 链接。此前文档站承诺每个模型都有 recipe 页, 裸脚本链接破坏了这一契约。

实现拆解

实现过程按以下步骤展开:

1. 新增三个 recipe 页面: docs/models/kimi/kimi-k3.md (176 行)、docs/models/nemotron/nemotron-3-ultra.md (172 行)、docs/models/gemma/gemma-4.md (151 行), 统一采用六段式结构(模型介绍、支持变体、环境准备、启动、配方配置、搭配阅读)。内容均从 scripts/run_nemotron_3_ultra_550b_a55b.py、scripts/run_gemma_4_26b_a4b.py、scripts/run_gemma_4_31b.py、scripts/models/* 及 launcher 快照中提取, 包括并行配置、SGLang 引擎参数、LoRA 目标模块、健康运行指标等。
2. 新增 / 更新家族索引: 新建 docs/models/gemma/index.md (39 行), 作为 Gemma 家族入口; 在 docs/models/nemotron/index.md 表格中加入 Ultra 行, 并补充“最大、跨 16 节点 latent MoE”的选型提示。
3. 更新导航配置: docs/docs.json 中 Kimi 分组加入 models/kimi/kimi-k3, Nemotron 分组加入 models/nemotron/nemotron-3-ultra, 并新建 Gemma 分组 (root 指向 models/gemma/index)。

4. 对齐两张模型总表：docs/index.md 与 docs/models/index.md 将 Kimi-K3、Nemotron-3-Ultra、Gemma-4 的链接从裸脚本 /PR 链接替换为 recipe 页链接，并补齐 Inkling-Small 与 JoyAI-LLM-Flash 行，使两份列表在家族顺序与条目上保持一致。
5. 清理过时指针：docs/models/thinkingmachines/inkling.md 删除已合并 PR#1683 的指引行。
6. 配套与修正：纯文档变更，无源码 / 测试配套；合入前由 maintainer Shi-Dong 在最终 commit 中修正了 Ultra 的多节点启动命令与 Gemma-4 31B 依赖的 Megatron-Bridge 分支名两处实质性错误。

关键文件：

- docs/models/kimi/kimi-k3.md（模块 模型文档；类别 docs；类型 documentation）：新增 Kimi-K3 LoRA RL 配方页，176 行，是三个新页面中信息量最大、技术细节最深的：覆盖 MXFP4 到 BF16 转换、32-rank torch_dist 转换、16 节点启动、LoRA 目标模块与健康运行指标，并明确标注内容来自未合入的 #1825 分支。
- docs/models/nemotron/nemotron-3-ultra.md（模块 模型文档；类别 docs；类型 documentation）：新增 Nemotron-3-Ultra-550B-A55B 配方页，172 行，核心价值在于把『Mamba n_groups=8 限制 TP 不超过 8』这条硬约束如何推导出 PP=4、EP=32 的完整布局讲透，并覆盖 4 层单节点切片与 16 节点全量两种变体。
- docs/models/gemma/gemma-4.md（模块 模型文档；类别 docs；类型 documentation）：新增 Gemma-4 配方页，151 行，覆盖 26B-A4B（MoE）与 31B（dense）两种形态，详细说明 sglang 保守内核选择与 routing replay 的必要性，并指出 31B 依赖 zhichen/gemma4-dense 分支。
- docs/models/gemma/index.md（模块 模型文档；类别 docs；类型 documentation）：新建 Gemma 家族索引页，作为 docs.json 中 Gemma 导航组的落地页，提供两张表格与选型建议，是导航结构变更的必要配套。
- docs/docs.json（模块 导航配置；类别 config；类型 configuration）：文档站导航配置，Kimi 组新增 kimi-k3、Nemotron 组新增 nemotron-3-ultra、新建 Gemma 组，是三个新页面能被导航发现的关键配置变更。
- docs/models/index.md（模块 模型文档；类别 docs；类型 documentation）：支持模型总表，新增 Kimi-K3 与 Nemotron-3-Ultra 行、补齐 Inkling-Small 与 JoyAI-LLM-Flash 行，同时与首页表格对齐，是本 PR 索引同步的核心文件之一。
- docs/index.md（模块 模型文档；类别 docs；类型 documentation）：首页模型表将 Kimi-K3、Nemotron-3-Ultra、Gemma-4 的裸脚本 /PR 链接替换为 recipe 页链接，是本次变更对最终用户最直接的呈现。
- docs/models/nemotron/index.md（模块 模型文档；类别 docs；类型 documentation）：Nemotron 家族索引补充 Ultra 行与选型建议，与新增 recipe 页配套。
- docs/models/thinkingmachines/inkling.md（模块 模型文档；类别 docs；类型 documentation）：删除已合并 PR#1683 的指针，保持『文档只引用已存在资源』的约定。

关键符号：未识别

评论区精华

该 PR 没有任何 review 评论，Shi-Dong 直接 APPROVE。唯一值得关注的是合入前最后一个 commit ([a2147fa](#)) 由 maintainer 直接修正的两处内容错误：Ultra launcher 没有 worker 子命令，多节点运行必须起外部 ray 集群并设置 `MILES_SCRIPT_EXTERNAL_RAY=1`；Gemma-4 31B 依赖的 Megatron-Bridge 分支实际是 `zhichen/gemma4-dense` 而非 `gemma4-dense`。这些修正表明合并阶段对“文档与脚本一致性”做了实质校验。

- Nemotron-3-Ultra 启动命令与 Gemma bridge 分支名勘误 (correctness): 两处错误已在 PR 内修复，文档与脚本行为对齐后再合并。

风险与影响

- 风险：纯文档变更，无运行时代码风险，但存在以下具体风险：
- 内容依赖未合入 PR：`docs/models/kimi/kimi-k3.md` 整页引用的脚本、`radixark/miles:kimi-k3` 镜像及 `sglang-miles-k3` 分支均来自未合入的 #1825；若读者在 main 上按文档操作会直接失败。页面虽已声明，但链接可解析性无法保证。
- 文档与脚本手工同步漂移：页面参数（如 `--rollout-max-concurrency` 默认值、`--sglang-mem-fraction-static`）由人从 launch 脚本提取，脚本后续变更不会自动同步到文档，存在再次漂移的可能。
- 无自动化校验：PR 检查仅人工核对链接与脚本，仓库没有针对 docs 链接的 CI 检查（如 `dead-link` 或 `libcst`），合并冲突（如与 #2391 在 `docs/models/index.md` 的行冲突）需要人工处理。
- 桥梁分支依赖：Gemma-4 31B 依赖 `zhichen/gemma4-dense` 个人分支，分支被删除或改名前文档即失效。
- 影响：影响范围集中在 docs/ 下 9 个文件：为三个模型家族补全了结构化配方文档，用户可以从裸脚本链接升级为带约束解释、变体对比、健康运行指标的教程；Kimi-K3 的“MXFP4 直训 + LoRA + colocated SGLang”16 节点配方首次获得完整说明。对团队而言，恢复了“每个模型都有 recipe 页”的文档契约，并让首页与模型总表呈现同一份模型清单，降低用户困惑与支持成本。对源码、运行时、构建链路均无影响。
- 风险标记：依赖未合入 PR#1825 的分支内容，文档与脚本手工同步易漂移，外部分支引用对个人仓库有依赖，无自动化链接校验

关联脉络

- PR #1825 Kimi-K3 LoRA RL implementation: Kimi-K3 配方页整页引用的脚本、镜像与 SGLang 分支均来自该未合入 PR；本 PR 文档的准确性依赖 #1825 最终落地。
- PR #2391 docs: replace DeepSeek V3/R1 page with a DeepSeek-V3.2 recipe: 与本 PR 在 `docs/models/index.md` 上产生行级冲突，合并时 DeepSeek 行取自 main，处理过程记录在提交 `09db190` 中。
- PR #2377 fix(kimi): align YaRN parameters with checkpoints: 同期 Kimi 系列文档与脚本对齐工作，与本 PR 同属让文档跟随 `scripts/` 实际内容的演进方向。
- PR #2216 docs: add GLM-5.2 model page and update supported-models tables: 同类变更：新增模型 recipe 页并同步 `docs.json` 导航与模型总表，是本 PR 的模式参照。