

PR #2594 完整报告

radixark/miles

docs: name the dense variant in the Qwen3.8 model-table rows

合并时间: 2026-08-18 13:11

原文链接: <http://prhub.com.cn/radixark/miles/pull/2594>

执行摘要

这是一个纯文档 PR，修改了 `docs/index.md` 与 `docs/models/index.md` 两个文件（合计 +4/-4）。核心动作是把模型速览表中的裸 `Qwen3.8` 统一改名为 `Qwen3.8-27B`，并在首页表格补上此前缺失的 `Qwen3.8` 行，同时扩展表格下方的指引文字提示完整家族列表位置。改动旨在消除“裸 `Qwen3.8` 同时承诺 `dense` 与 `sparse` 两个成熟度不同变体”的歧义，不对任何代码、测试或导航结构产生影响。

功能与动机

PR body 明确指出，`Qwen3.8 recipe` 页面覆盖两个成熟度差异很大的模型：

变体	状态
Qwen3.8-27B (dense)	已合并， <code>python scripts/run_qwen3_dense.py --model-name Qwen3.8-27B</code> ，实测约 12 min/step (1 x 8 H200)
Qwen3.8-2.4T-A95B (sparse)	launcher 仍在 #2488 中，依赖 LoRA 分支，仅 LoRA 可用

因此“裸 `Qwen3.8` 行读起来像是对两个变体的共同承诺”，而 `Qwen3.8-27B` 的命名方式与兄弟行（`Qwen3.5-4B / 9B / 27B`、`Qwen3.6-27B`）保持一致。另外，首页表格自称“is not exhaustive”却从未说明遗漏模型去向（GLM-4.5、Kimi-K2、Qwen3 / Qwen3-Next、gpt-oss-20b 等都在 `/models` 页面），本次用一句话闭环。

实现拆解

- 修正 `/models` 索引表格 (`docs/models/index.md`) - 将 `Qwen` 行首条链接文本 `[Qwen3.8](/models/qwen/qwen3-8)` 改为 `[Qwen3.8-27B](/models/qwen/qwen3-8)`。 - 链接目标不变，仅文本变更；目的是避免对 `sparse` 变体的隐含承诺，并与其他带尺寸的兄弟命名对齐。
- 补齐首页速览表格 (`docs/index.md`) - `Qwen` 行新增 `[Qwen3.8-27B](/models/qwen/qwen3-8)` 并置于行首。 - 补上前情：#2575 只在 `/models` 页面添加了 `Qwen3.8` 记录，首页从未有对应行；本次让两个文档入口保持一致。
- 扩展首页表格下方指引文字 - 由“exact conversion commands, launch scripts, and parallelism settings”扩展为“the full family lists, exact conversion commands, launch

scripts, and parallelism settings”。 - 目的：告知读者首页未穷尽的模型都存在 [/models](#) 索引页。

4. 无配套变更 - 未涉及代码、测试、配置或部署配套改动，PR body 明确声明“Docs-only; no code, tests or navigation touched.”。

以下是两个文档文件的变更后片段（注释追加说明）。

```
<!-- docs/index.md 首页模型速览表：Qwen 行新增 Qwen3.8-27B 条目并置于行首，此前该行缺失 -->
| **Qwen** | [Qwen3.8-27B](/models/qwen/qwen3-8)<br/>[Qwen3.6 MoE](/models/qwen/qwen3-6-moe)<br/>[Qwen3.6](/models/qwen/qwen3-6)<br/>[Qwen3.5-35B-A3B](/models/qwen/qwen3-5-moe)<br/>[Qwen3.5-4B / 9B / 27B](/models/qwen/qwen3-5) |
```

```
<!-- 表格下方指引文字：补上“完整家族列表”的表述，告知首页未穷尽的模型在 /models 索引页 -->
See [Models](/models/index) for the full family lists, exact conversion commands, launch scripts, and parallelism settings.
```

```
<!-- docs/models/index.md: Qwen 行首条由裸 Qwen3.8 改为 Qwen3.8-27B，与兄弟行命名对齐 -->
| **Qwen** | [Qwen3.8-27B](/models/qwen/qwen3-8)<br/>[Qwen3.6 MoE](/models/qwen/qwen3-6-moe)<br/>[Qwen3.6](/models/qwen/qwen3-6)<br/>[Qwen3.5-35B-A3B](/models/qwen/qwen3-5-moe)<br/>[Qwen3.5-4B / 9B / 27B](/models/qwen/qwen3-5)<br/>[Qwen3-Next-80B-A3B-Thinking](/models/qwen/qwen3-next)<br/>[Qwen3-30B-A3B / 235B-A22B](/models/qwen/qwen3-moe)<br/>[Qwen3-0.6B / 1.7B / 4B / 8B / 14B / 32B](/models/qwen/qwen3) |
```

评论区精华

本 PR 没有实质技术讨论。claude[bot] 自动评论提示“This repository is configured for manual code reviews. Comment [@claude review](#) for a one-time review”，随后 Shi-Dong 直接批准（APPROVED），无任何 inline 评论。所有决策依据都集中在 PR body 与 commit message 中，即对“裸命名会同时承诺两个成熟度不同变体”的主动修正。

风险与影响

- 文档与实现状态耦合：Qwen3.8-27B 的命名与“dense 已合并、sparse 仍在 #2488”这一状态绑定。一旦 #2488 合并，sparse 变体具备可运行性后，需再次更新两个表格，否则文档会落后于实现。
- 链接目标单一：两个表格的链接共同指向 [/models/qwen/qwen3-8](#) 单一 recipe 页，该页内部是否分别标注两个变体的成熟度不在本次改动范围内，歧义只是从表格层转移到页面层。
- 影响评估：仅影响文档读者；对用户价值在于能快速区分可运行的 dense 变体与未发布的 LoRA-only sparse 变体，避免误用。无代码回归、性能或安全风险。

关联脉络

本 PR 是 #2575（新增 Qwen3.8-27B recipe）的后续补丁：#2575 只将 [Qwen3.8](#) 加到了 [/models](#) 索引页，本 PR 补上首页入口并统一命名。同时，#2488（Qwen3.8 sparse 变体 launcher）被 PR body 引用为改名的直接动机。整体来看，这是 Qwen3.8 模型支持进入文档收尾阶段的细节完善，未来的触发点是 #2488 合并后对两个表格的再次同步。