# PR #2575 完整报告

radixark/miles

feat: add the Qwen3.8-27B recipe

合并时间：2026-08-17 05:50

原文链接：http://prhub.com.cn/radixark/miles/pull/2575

## 执行摘要

- 一句话：新增 Qwen3.8-27B 训练 recipe，复用 Qwen3.5 模型参数
- 推荐动作：建议精读 docs/models/qwen/qwen3-8.md 第 5.3 节，它把“为什么把 mem-fraction 从 0.5 调到 0.8”讲得很完整（GDN 状态 151 MB/ 序列、KV pool 对比、吞吐对比），是很好的性能调优案例。代码层面只有一行 recipe 和一行参数派生，含金量不高，但“同构模型复用参数契约 + 快照固化展开结果”的维护策略值得作为新增模型的标准模板。

## 功能与动机

Qwen3.8 是 Qwen3 系列的新版本，miles 需要为 dense 27B 提供开箱即用的 RL 训练入口。关键事实是 Qwen3.8-27B 与 Qwen3.5-27B 完全同构——同样 64 层、hidden-size 5120、vocab 248320、GDN 混合注意力骨架——连 config.json 都逐字节一致，因此不需要新的 Megatron spec，只需挂到现有 dense 训练路径上。文档第 6 节还专门说明 2.4T-A95B 稀疏版是另一套 NVFP4/BF16 混合的 LoRA 方案，不在本页 launcher 内，避免用户误用。

## 实现拆解

1. 模型参数契约：新增 scripts/models/qwen3.8-27B.py（6 行），定义 model_args() -> str 并委托 load_sibling_model_args(__file__, "qwen3.5-27B") 生成整条 Megatron flag。这样 Qwen3.5-27B / Qwen3.6-27B / Qwen3.8-27B 三者的展开结果字节级一致，避免复制一份易漂移的参数列表。
2. 训练入口扩展：scripts/run_qwen3_dense.py 中 _MODEL_NAMES 的 Literal 类型增加 "Qwen3.8-27B"，随后在 _RECIPES 注册表新增 _Recipe("qwen3.8-27B", 4, 8192, 1, 0.8, True)。五个位置参数依次是 Megatron 模型类型、TP 大小、每 GPU token 预算、每引擎 GPU 数、SGLang mem-fraction；最后布尔值表示开启 CPU Adam offload。与 Qwen3.5-27B 的唯一差异是 mem-fraction 从 0.5 提到 0.8，其余继承。
3. 调参依据：docs 第 5.3 节解释 0.8 的由来——该模型 48/64 层为线性注意力，SGLang 侧 per-sequence GDN 状态约 151 MB，0.5 时 KV pool 仅 158k token、引擎并发 11-12、decode 553-690 tok/s；0.8 时 KV pool 518k、并发 39、decode 1855 tok/s，训练峰值 88-90 GB 未超 140 GB。文档还给出两个后续杠杆：--rollout-num-gpus-per-engine 2 或 --sglang-mamba-ssm-dtype bfloat16。
4. 文档与导航配套：新增 docs/models/qwen/qwen3-8.md（193 行），包含模型简介、支持变体表、HF 下载与 convert_hf_to_torch_dist 转换命令、quick-start、每步成本实测表、

并行 / 算法 /rollout/optimizer 配置、notable quirks、2.4T-A95B 说明。docs/docs.json 在 Qwen 分组下新增 Qwen3.8 折叠组并置顶；docs/models/qwen/index.md 补齐 Qwen3.6 dense/MoE 与 Qwen3.8 三行；docs/models/index.md 的 Qwen 链接列表头部加入 Qwen3.8。

5. 快照与测试配套：新增 tests/snapshots/model_args/qwen3.8-27B.txt，固化 model_args() 展开后的 34 个 flag（--spec miles_plugins.models.qwen3_5 get_qwen3_5_spec、--rotary-base 10000000、--rotary-percent 0.25、--vocab-size 248320、--attention-output-gate 等）。本次未新增测试文件，快照供既有的模型参数一致性测试比对；CI 策略无改动。

关键文件：

- scripts/models/qwen3.8-27B.py（模块 模型脚本；类别 source；类型 data-contract；符号 model_args）：模型参数契约入口：通过 load_sibling_model_args 复用 Qwen3.5-27B 的完整 Megatron flag 定义，是本次模型支持的核心机制。
- scripts/run_qwen3_dense.py（模块 训练入口；类别 source；类型 core-logic）：dense 训练入口：_RECIPES 注册表新增 Qwen3.8-27B 条目，mem-fraction 从 0.5 提到 0.8，是唯一的行为差异。
- docs/models/qwen/qwen3-8.md（模块 模型文档；类别 docs；类型 documentation）：核心交付物：193 行 recipe 文档，含架构说明、转换 / 启动命令、实测成本表、mem-fraction 0.8 的完整论证与 2.4T-A95B 预告。
- tests/snapshots/model_args/qwen3.8-27B.txt（模块 参数快照；类别 docs；类型 documentation）：参数快照：固化 model_args() 展开后的 34 个 Megatron flag，供参数一致性测试比对，防止复用派生漂移。
- docs/docs.json（模块 文档导航；类别 config；类型 configuration）：文档导航：Qwen 分组下新增 Qwen3.8 折叠组，使新页面进入侧边栏。
- docs/models/qwen/index.md（模块 模型索引；类别 docs；类型 documentation）：Qwen 模型索引：补上 Qwen3.6 dense/MoE 与 Qwen3.8 三行，保持模型总表完整。
- docs/models/index.md（模块 模型索引；类别 docs；类型 documentation）：全模型导航列表：Qwen 条目头部加入 Qwen3.8 链接。

关键符号：model_args

## 关键源码片段

### scripts/run_qwen3_dense.py

dense 训练入口：_RECIPES 注册表新增 Qwen3.8-27B 条目，mem-fraction 从 0.5 提到 0.8，是唯一的行为差异。

```
# Qwen3.8-27B 与 Qwen3.5/Qwen3.6 共用同一套 dense 训练路径，
# 差异只体现在 _Recipe 的 knob 上（TP、token 预算、mem-fraction、offload 等）。

_RECIPES: dict[str, _Recipe] = {
    "Qwen3-4B": _Recipe("qwen3-4B", 2, 9216, 2, 0.7, False, use_dashboard=True),
    # SGLang 0.5.9 下 TP>1 对 Qwen3.5 输出异常，miles 仍锁该版本，
```

```
    # 因而整条 Qwen3.5+ 线每引擎只分配 1 GPU（见 sglang issue #21039）。
    "Qwen3.5-4B": _Recipe("qwen3.5-4B", 2, 9216, 1, 0.7, False),
    "Qwen3.5-9B": _Recipe("qwen3.5-9B", 2, 9216, 1, 0.6, False),
    "Qwen3.5-27B": _Recipe("qwen3.5-27B", 4, 8192, 1, 0.5, True),
    "Qwen3.6-27B": _Recipe("qwen3.6-27B", 4, 8192, 1, 0.5, True),
    # Qwen3.8-27B 与 Qwen3.5-27B 同构，唯一差异是把 SGLang mem-fraction
    # 从 0.5 抬到 0.8：48/64 层线性注意力的 GDN 状态约 151 MB/ 序列，
    # 0.5 时 KV pool 仅 158k token，0.8 后 decode 吞吐从 ~600 升到 1855 tok/s。
    "Qwen3.8-27B": _Recipe("qwen3.8-27B", 4, 8192, 1, 0.8, True),
}
```

## 评论区精华

本 PR 的 review 周期非常简短：claude[bot] 发出自动提示，说明仓库配置为人工 code review，可通过 @claude review 触发一次性或持续 review；guapisolo 直接 APPROVED，无评论内容。没有产生任何 inline review comments，技术决策（mem-fraction 0.8）由作者在文档中完成论证，未经历讨论回合。

- 人工 review 流程提示 (other): 无实质技术讨论；guapisolo 最终直接 APPROVED，无评论内容。

## 风险与影响

- 风险：
  1. spec 复用级联影响：scripts/models/qwen3.8-27B.py 的 model_args() 完全委托 qwen3.5-27B，未来 Qwen3.5 参数定义变更（如 rotary-base、vocab 调整）会静默传导给 Qwen3.8，需在修改 qwen3.5-27B 时注意快照测试。
  2. SGLang 版本锁定：miles 仍锁 SGLang 0.5.9，已知 TP>1 对 Qwen3.5 系列生成异常（issue 21039），Qwen3.8-27B 作为同构模型同样受限，recipe 强制每引擎 1 GPU。
  3. mem-fraction 硬件依赖：0.8 是作者在 H200 140 GB 上实测的调优值（训练峰值 88-90 GB）；换成 A100 80G 等低显存环境未验证，直接照搬 recipe 可能 OOM。
  4. 测试覆盖空缺：本次只有快照数据，没有新增单测或 E2E 覆盖 Qwen3.8-27B 启动路径，回归主要靠现有 dense 路径测试兜底。 - 影响：对用户：多了一个受支持模型，--model-name Qwen3.8-27B 可直接启动 GRPO 训练；文档包含下载、转换、启动、成本表与调优建议，显著降低上手成本。对系统：训练侧无核心逻辑改动，只是配置矩阵增加一行；SGLang 推理侧因 mem-fraction 0.8 显存水位更高、并发更大。对团队：Qwen 系列文档补齐了 Qwen3.6 缺失行；文档提前预告 #2488 的 2.4T-A95B 方案，避免用户误用 dense recipe 跑 MoE 模型。 - 风险标记：spec 复用级联影响，SGLang 版本锁定，mem-fraction 硬件依赖，缺少新增测试

## 关联脉络

- PR #2488 Qwen 3.8 day-0 lora RL support: PR #2575 文档第 6 节明确引用 #2488：Qwen3.8-2.4T-A95B 的 launcher（scripts/run_qwen3_8.py）与 NVFP4 LoRA recipe 在 #2488 落地，与本 PR 构成 Qwen3.8 系列支持的两半。