PR #2650 完整报告

radixark/miles

docs: flesh out the Qwen3.8-2.4T-A95B recipe on the Qwen3.8 page

合并时间：2026-08-18 20:06

原文链接：http://prhub.com.cn/radixark/miles/pull/2650

执行摘要

- 一句话：Qwen3.8-2.4T 训练 recipe 文档化，替换 PR 占位 stub
- 推荐动作：对计划在 GB300 上跑 Qwen3.8-2.4T-A95B LoRA RL 的工程师：值得精读，§6.2 的 NCCL 版本坑与 §6.4 的交替备份内存预算属于实测得出的硬经验。对文档维护者：注意本文档与 #2488 的强耦合，建议在 #2488 合并时同步 review 本文档。值得关注的设计决策是「任意时刻恰好一边持有 host backup」的内存预算推演，以及用 zero_copy=True 避免 clone 备份膨胀 malloc arena 的做法。

功能与动机

原 §6 是 PR 指针 stub，明确写着「That recipe is not part of this page's launcher. It lands separately in #2488」且「PR #2488 is still open … the recipe and its reproduction steps may still move」。PR body 说明本次目标是「Replaces the PR-pointer stub with the validated recipe」，按 #2488 launcher 接口和 Kimi-K3 页面风格补齐变体、环境、启动、内存四小节，并给出 GSM8K 0.639→0.75、DAPO eval/aime 0.633→0.933 的验证结果作为采用依据。

实现拆解

变更入口：docs/models/qwen/qwen3-8.md 是本次唯一改动文件，将 §6 Qwen3.8-2.4T-A95B 从「仅指向 #2488 的占位说明」重写为可直接复现的完整 recipe（+97/-17）。

步骤 1：定义变体（§6.1） 新增 Variants 小节，说明 --model-variant 的两种取值：4layer（4 层 smoke，默认，单节点 8 GPU）与 full（92 层完整模型，16 × 4 GB300 共 64 GPU，已验证）。变体同时决定 checkpoint 路径与 megatron_model_type，对应 scripts/models/qwen3.8-2.4T-A95B{_4layer,_full}.py 两个定义文件。这里把「变体」作为 recipe 的第一个决策点，让读者先明确自己是在 smoke 还是跑完整模型。

步骤 2：环境准备（§6.2） 规定统一使用多架构镜像 docker.io/radixark/miles:dev（同一 tag 覆盖 GB300 aarch64 与 x86 节点）。重点记录 NCCL 版本坑：多节点 rollout 引擎要求 NCCL ≥ 2.30.7，2.28.x 会在 CUDA-graph 捕获的跨节点集合通信上死锁，并给出 nvidia-nccl-cu13==2.30.7 安装与 SGLANG_NCCL_SO_PATH/LD_PRELOAD 注入命令。随后说明 launcher 的 checkpoint 查找规则（--model-dir 下 Qwen3.8-2.4T-A95B-NVFP4_<variant> 与 Qwen3.8-2.4T-A95B-bf16_<variant>）以及内置

prepare 步骤：下载 zhuzilin/dapo-math-17k、执行 BF16 → torch_dist 转换（tools/convert_hf_to_torch_dist.py），转换布局与训练布局解耦（加载时重新分片）。

步骤 3：启动命令（§6.3） 给出两条命令：单节点默认 smoke（python scripts/run_qwen3_8.py）与 16 × 4 GB300 全量命令。全量命令的关键参数包括 trainer TP1/PP4/EP16/ETP1、rollout 4 引擎 × TP16、LoRA rank 32/alpha 64 仅作用于 attention 投影（q,k,v,o_proj），并说明 num_query_groups 4 对 native-LoRA TP 上限的影响，以及 69 个 GDN mixer 层与 routed experts 无 adapter、rollout MoE 路径完全不变的理由。

步骤 4：内存预算（§6.4） 这是全篇工程含量最高的小节：956 GB 主机的可行前提是「任意时刻恰好一边保留 host backup」——rollout 期间 sleeping trainer 的 tms 备份约 390 GB/node，训练期间 sleeping engines 的 NVFP4 镜像约 372 GB/node（由 --lora-base-cpu-backup 开启，同时让每次权重更新跳过 base 重发）。随后给出三条保底线 flag：--drop-checkpoint-page-cache-after-load（防多 TB DCP 读取的 page cache 与 pinned 分配峰值竞争）、--colocate-memory-peak-device gpu 与禁用 --use-kl-loss（零系数仍会加载完整 ref checkpoint）、权重同步必须使用 get_cpu_backup(zero_copy=True)（clone 变体每次更新保留约 75 GB/rank malloc arena）。健康峰值 600–650 GB/node，持续超 750 GB 即说明第二份备份或 page cache 回归。

步骤 5：验证数据与配套 无源码与测试联动。PR body 与 commit message 提供验证结果：GSM8K eval 0.639 → 0.75（40 steps）、DAPO eval/aime 0.633 → 0.933（step 100）。commit message 还提到转换规则涉及 fused-expert EP offset 与逐比特产物审计、rollout 引擎池按 workload-pinned 方式分配，但正文窗口未完整展开这些细节，需要回查 #2488。

关键文件：

- docs/models/qwen/qwen3-8.md（模块 模型文档；类别 docs；类型 documentation）：唯一变更文件，将 §6 Qwen3.8-2.4T-A95B 从 PR 占位 stub 重写为经过验证的完整训练 recipe，包含变体、环境、启动、内存四小节及验证数据。

关键符号：未识别

评论区精华

本 PR 没有实质评审交锋：claude[bot] 仅回复仓库已配置人工 review 的通知，guapisolo 直接 APPROVED 且未留文字。工程层面的关键判断全部集中在 PR body 与 commit message 自述中，例如 commit message 提到转换规则需审计 fused-expert EP offset 与逐比特产物、rollout 引擎池按 workload-pinned 方式分配、交替 host-backup 内存预算等，但这些细节在文档正文中未逐条展开，具有不可忽略的置信度风险。

- 暂无高价值评论线程

风险与影响

- 风险：文档与 #2488 实现强耦合：本文档描述的 launcher 接口、模型 args、内存 flag（如 get_cpu_backup(zero_copy=True)、--lora-base-cpu-backup）全部来自 #2488，若 #2488 演进或合并时参数变化，本文档会迅速失真。内存预算数值（390/372 GB、600–650 GB 峰值、75 GB/rank）基于 16 × 4 GB300 实测，其他拓扑、机型或并行配置

需重新标定，直接套用可能误导用户。NCCL 2.30.7 的注入路径 /usr/local/lib/python3.12/dist-packages/nvidia/nccl/lib/libnccl.so.2 依赖镜像内部 Python 版本与包布局，镜像更新可能导致路径失效。无运行时源码变更，因此无直接回归风险，但命令中的 --tp 1 --pp 4 --ep 16 等参数若与 #2488 实际实现不一致，会产生复现层面的误导。

- 影响：对用户（文档读者）：从「只能追 PR 等变动」升级为「可按文档独立复现」，显著降低 Qwen3.8-2.4T-A95B 后训练的上手成本，尤其 NCCL 版本坑和内存预算推演能帮用户避开两类最隐蔽的失败模式。对团队：文档化的验证数据（GSM8K 0.639→0.75、DAPO AIME 0.633→0.933）为后续稀疏 MoE 模型的 recipe 文档提供了基线与参照风格。对系统：无任何运行时代码影响。影响范围集中在文档与模型复现链路，影响程度中低。
- 风险标记：文档依赖 #2488 未合入实现，内存数值依赖特定硬件验证，NCCL 路径依赖镜像内部结构

关联脉络

- PR #2488 Qwen 3.8 day-0 lora RL support: 本文档描述的 launcher（scripts/run_qwen3_8.py）与模型定义（scripts/models/qwen3.8-2.4T-A95B{_4layer,_full}.py）全部来自该 PR，是 recipe 的实现落地，两者强耦合。
- PR #2575 feat: add the Qwen3.8-27B recipe: 同一 Qwen3.8 页面上的 dense 27B recipe，共享页面结构、启动脚本风格与文档组织方式。
- PR #2594 docs: name the dense variant in the Qwen3.8 model-table rows: 同属 Qwen3.8 文档清理线，将裸 Qwen3.8 命名改为 Qwen3.8-27B，与本文档的稀疏 / 稠密区分逻辑一致。