PR #2478 完整报告

radixark/miles

docs: rewrite INT4 QAT guide

合并时间：2026-08-13 04:13

原文链接：http://prhub.com.cn/radixark/miles/pull/2478

执行摘要

本 PR 重写 docs/advanced/int4-qat.md，把 INT4 QAT 从旧的“通用 W4A16 训练内存路径”描述修正为 miles 的真实实现：Megatron 以训练 dtype 保存可训练权重，仅对 routed-expert 的 Transformer Engine GroupedLinear 在 forward 中做对称 INT4 fake 量化，SGLang 侧加载 compressed-tensors packed W4A16 rollout 权重。文档新增组件职责表、权重生命周期、量化设置与 Kimi-K2.5/Qwen 两条路径的说明，并删除不支持的 maturity、memory-saving 等声明，属于一次高质量的实现对齐型文档重写。

功能与动机

PR body 明确说明：旧页面把 INT4 QAT 描述为通用 W4A16 训练内存路径，并以旧 shell 配方为中心，与实际实现不符。实际实现更窄：Megatron 保留配置 dtype 的可训练权重，在 Transformer Engine GroupedLinear 中做对称 fake INT4 量化，并导出 packed routed-expert 张量供 SGLang rollout。重写的动机就是把流程各部分分配给真实所属组件，区分 rollout checkpoint 的 compressed-tensors 存储契约与 Megatron 独立配置的 fake-QAT group size，同时移除不被支持的功能宣称。

实现拆解

1. 结构重写与定位修正：更新文档标题描述，将 INT4 QAT 定位为“MoE 策略在 Megatron 中以 fake 量化专家权重训练、SGLang 以 packed W4A16 rollout 权重服务”，新增组件职责表，拆分 Hugging Face checkpoint、Megatron、Megatron Bridge/raw-mode exporter、SGLang 四类角色。
2. 补充权重生命周期与量化公式：在 docs/advanced/int4-qat.md 中新增 5 步 weight lifecycle，从 --hf-checkpoint 初始化、Megatron 训练 dtype 初始化、fake 量化 forward（STE 反向）、权重导出打包、到 SGLang weight-update session 加载；同时给出对称 INT4 fake 量化公式。
3. 明确 quantization settings 与配置独立性：记录 rollout 侧 compressed-tensors pack-quantized 契约（num_bits: 4、strategy: group、symmetric: true、仅打包 GroupedLinear 覆盖的 routed expert projections、末维需能被 group size 整除）；强调训练侧 OPEN_TRAINING_INT4_GROUP_SIZE 与 checkpoint group size 是两个独立实现，匹配不等于 bitwise 一致，改变环境变量不转换 checkpoint。
4. 补充 Kimi-K2.5 与 Qwen 路径：记录 Kimi Bridge 双向转换（HF→Megatron 解包 INT4 到 BF16，Megatron→HF 重新打包为 group-size-32 INT4 并返回 weight_packed/weight_scale/weight_shape）；说明 --ref-load BF16 只是当前

launcher 写法而非 Bridge 要求；记录 Qwen 路径 rollout packing 用 group size 32、trainer fake QAT 用 128 的现状。

5. 删除不支持声明：移除关于通用内存节省、成熟度、兼容性及旧 workflow 的过度宣称。

该 PR 为纯文档变更，无源码、测试或配置联动；提交历史显示作者在 5 次提交中逐步完成结构重写、Bridge 精度澄清、数据路径对齐、声明收敛与最终审计。

本 PR 无源码改动，以下为重写后文档中最值得关注的核心内容整理：

miles 将 INT4 QAT 拆分为四个组件职责：--hf-checkpoint 定义 rollout 存储格式；Megatron 保持权重为训练 dtype 并在 forward 前对 routed-expert GroupedLinear 做对称 fake 量化；Megatron Bridge 或 raw-mode exporter 负责 HF 命名映射与打包；SGLang 以 compressed-tensors W4A16 契约加载。

对权重组 w 的 fake 量化公式：

```
scale  = max(max(abs(w)) / 7, 1e-5)
fake_w = clamp(round(w / scale), -7, 7) * scale
```

注意：fake 量化只在 forward 中模拟 INT4，训练权重、优化器状态、梯度与激活仍保持配置 dtype；训练侧 OPEN_TRAINING_INT4_GROUP_SIZE 与 rollout 的 group size 是两个独立配置项，匹配二者不等于 bitwise 一致。

评论区精华

该 PR 没有产生 review 评论或讨论线程，最终由 Zhichenzzz 直接批准。核心质量保障体现在 PR body 的 validation 说明中：作者对照 miles 实现、Megatron-LM miles-main、Docker-pinned Megatron-Bridge commit 以及 SGLang origin/sglang-miles 逐一审计文档声明，并完成 git diff --check、pre-commit、内链检查与代码围栏平衡校验。

风险与影响

1. 文档与实现漂移风险：文档精确描述了当前 pinned Megatron Bridge、Transformer Engine GroupedLinear 行为以及 SGLang packed W4A16 契约，一旦相关组件升级或改用其他量化实现（如离线 GPTQ 转换），文档可能迅速过期。
2. 配置误导风险：文档明确指出 OPEN_TRAINING_INT4_GROUP_SIZE 与 checkpoint group size 相互独立、且改变环境变量不会转换 checkpoint；若读者忽略该提示，仍可能误以为修改环境变量即可完成量化迁移。
3. 行为收窄的认知落差：移除 memory-saving 与通用 W4A16 路径的描述后，期望利用 INT4 QAT 降低显存 / 存储的用户需要重新评估预期；文档已经说明不降低训练权重、优化器、梯度与激活的存储，但这一信息需要被新读者及时看到。
4. 无自动化校验：文档中的公式、flag 与路径均为人工审计，后续重构缺少能自动捕获文档过期的手段。

影响对象主要是使用 INT4 QAT 的用户与参考该路径复现 Kimi-K2.5、Qwen3 INT4 配方的工程师：文档将“训练侧 fake 量化”与“rollout 侧 packed W4A16 存储格式”彻底分离，能显著减少对容量规划、权重导出与 group size 语义的误解。对团队而言，文档与真实实现的对齐降低了未来基于该指南做二次开发或问题排查的沟通成本。

关联脉络

该 PR 与近期多个文档与启动器演进直接相关：

- PR#2396 刷新 MXFP8/NVFP4 低精度 RL 指南，同属量化路径文档校准系列；
- PR#2356 将 .sh 启动脚本全部替换为 .py 启动器，本指南引用的 scripts/run_kimi_k25.py 正是该体系的一部分；
- PR#2375 重组 Training Backends 文档并提升 FSDP 地位，INT4 QAT 指南中的 Megatron Bridge 路径与后端能力描述在文档体系上相互衔接。

整体来看，miles 正在把低精度量化和训练后端的能力边界在文档中逐步收敛到真实实现，减少过度承诺，为后续基于这些路径的复现和产品化提供更可靠的依据。