

PR #2482 完整报告

radixark/miles

examples: fix READMEs — dead links, broken list structure, missing entries

合并时间: 2026-08-13 06:26

原文链接: <http://prhub.com.cn/radixark/miles/pull/2482>

执行摘要

- 一句话: 修复 examples 各 README 死链、列表结构与缺失条目
- 推荐动作: 本 PR 值得 5 分钟快速浏览 diff: 它是“文档内容为自动化铺路”的干净范例, 也是 stacked PR 顺序管理 (#2482 先行、#2481 重定向 main 后跟上) 的直观教材。对文档维护者, examples/README.md 作为 meta description 来源的写法值得借鉴。对工程师而言, 真正值得精读的是 #2481 的 scripts/tools/sync_example_docs.py 生成器——它把死链、H1、描述长度变成机器可检查的契约, 本 PR 只是这份契约的输入整理。

功能与动机

本 PR 是对 examples/** README 的一次内容审计修复。PR body 说明这些改动是在为 docs 站镜像做审计时发现的, 镜像机制 (#2481) 会单独落在本 PR 之上 ("found while auditing the READMEs for the docs-site mirror, which lands separately on top of this PR (#2481)")。修复对象包括: #1953 删除 examples/retool 后 retool_v2 README 仍指向旧目录; p2p_weight_transfer 的文档链接使用了从未存在的 docs/en/ 前缀且层级少一级; low_precision 的 Quick Start 编号步骤被无缩进的段落和代码块拆散; 顶层索引缺少 agentenv、nemo-gym、verifiers 等条目。关联 Issue #2481 还说明, 镜像生成器的 dead-link check、required-H1 rule 和 description-length gate 都会对未修复的 README 失败, 因此本 PR 必须先行合入。

实现拆解

变更入口

examples/README.md 是枢纽: 它既是 GitHub 上所有示例的导航入口, 也是 #2481 镜像生成器读取页面清单与 meta 描述的地方; 其余 4 个文件是审计中发现的独立问题点。

1. 重构顶层索引分组 (examples/README.md, +15/-11) 把 "Directory Structure" 的仓库布局视角, 改为三个面向读者的分组: Recipes (端到端训练配方)、Infra Features (运行期与基础设施管道)、Experimental (未完全验证), 每组配一行引导语; 补上 agentenv、nemo-gym、verifiers 三个缺失条目和 geo3k_vlm/multi_turn、on_policy_distillation/qwen3_5_35b_selfdistill 两个子配方。原因: 这些示例真实存在于仓库但无法被读者发现; 且 #2481 生成器以本文件条目描述作为镜像页面的 meta description, 条目不全意味着 docs 站 Examples 标签页缺页。

2. 清除对已删除内容的引用 (`examples/retool_v2/README.md`, +4/-2) 移除对 `examples/retool` 的链接与“参照 retool v1 README”的指引, 改为如实描述 `run_retool_multi_turn.py` 的自足行为: 脚本自行下载 dapo-math-17k 训练集与 aime-2024 评测集、下载 checkpoint 并在训练前转换为 `torch_dist`。这消除了读者被引向 404 的路径。
3. 修正文档链接 (`examples/infra_features/p2p_weight_transfer/README.md`, +1/-1) 链接从 `docs/en/advanced/p2p-weight-transfer.md` 改为 `./../docs/advanced/p2p-weight-transfer.md`。原链接同时犯了两个错误: `docs/en/` 前缀不存在于本仓库; 从示例子目录到仓库根需要三层 `../` 而非两层。
4. 修复列表结构 (`examples/infra_features/low_precision/README.md`, +21/-21) 把 Quick Start 每个编号步骤下的说明段落、代码块和注意事项全部缩进嵌套回对应步骤下, 并顺带把裸的三反引号围栏补全为带 `bash` 语言标记的 fenced code block。该文件在 GitHub 和 docs 站镜像下都要求结构规整。
5. 补齐标题与引言 (`examples/infra_features/README.md`, +6/-4) 为原本没有标题的索引页补上 `# Infra Features` H1, 并将两段式引言压缩为单段。H1 同时是 #2481 镜像生成器的必需项 (required-H1 rule)。

配套改动

无测试、配置、schema 或部署改动; PR 说明 pre-commit 全量通过, 并验证了所有被触碰 README 的链接都能在仓库树中解析。唯一需要跨 PR 配合的是合并顺序: 本 PR 先合入 main, #2481 随后重定向到 main。

关键文件:

- `examples/README.md` (模块 示例索引; 类别 docs; 类型 documentation): 顶层示例索引, 本次重构为 Recipes / Infra Features / Experimental 三个读者向分组, 补上 `agentenv`、`nemo-gym`、`verifiers` 三个缺失条目和 `multi_turn`、`qwen3_5_35b_selfdistill` 两个子配方; #2481 生成器会以这里的条目描述作为 docs 站页面 meta 来源, 是整批修复的枢纽。
- `examples/infra_features/low_precision/README.md` (模块 基础特性; 类别 docs; 类型 documentation): Quick Start 的编号步骤原先被无缩进的段落和代码块拆散, 在 GitHub 上每步都脱离自己的说明, 是本次修复中结构性问题最重的文件; 同时补全了代码围栏语言标记。
- `examples/retool_v2/README.md` (模块 工具调用; 类别 docs; 类型 documentation): 原 README 指向 `examples/retool` (#1953 已删除) 并让用户“去那边看数据准备”, 是典型的死链; 现在如实描述 `run_retool_multi_turn.py` 的自足行为。
- `examples/infra_features/p2p_weight_transfer/README.md` (模块 权重传输; 类别 docs; 类型 documentation): 文档链接用了从未存在的 `docs/en/` 前缀且少一层目录; 加上缺失的 `../` 后指向仓库真实文档, 是链接修正的最小但必要的改动。
- `examples/infra_features/README.md` (模块 基础特性; 类别 docs; 类型 documentation): 之前没有标题, 在 GitHub 上渲染出来只有一段话; 补上 H1 并压缩引言。H1 规则同时是 #2481 生成器的必需项。

关键符号: 未识别

关键源码片段

examples/infra_features/low_precision/README.md

Quick Start 的编号步骤原先被无缩进的段落和代码块拆散，在 GitHub 上每步都脱离自己的说明，是本次修复中结构性问题最重的文件；同时补全了代码围栏语言标记。

<!-- 修复要点：原版本第 1 步与第 2 步之间的段落和代码块没有缩进，在 GitHub 渲染时编号列表被拦腰打断，每步说明都与自己的步骤编号脱离；现在把每步的附属内容统一缩进嵌套在步骤下，同时把裸的嵌入式代码围栏补上 bash 语言标记 -->

Quick Start

1. Check if your training script is properly configured.

For training tasks, we need to add these flags:

```
```bash
--fp8-format e4m3
--fp8-recipe blockwise
--fp8-param-gather # [optional] Currently incompatible with CPU Adam
```
```

Then ensure the `NVTE\_FP8\_BLOCK\_SCALING\_FP32\_SCALES` environment variable is enabled.

Note that only `Linear` and `GroupLinear` layers in TransformerEngine use fp8 format. `embedding` and `lm\_head` remain in their original precision. If `--fp8-param-gather` is not enabled, weights in TransformerEngine remain in bf16, only being cast to fp8 during `GEMM` or `GroupGEMM` operations.

2. Convert your HuggingFace model weights to FP8 format.

You can use `tools/convert\_hf\_to\_fp8.py` to convert bf16 weights to fp8 format. Ensure the `--hf-checkpoint` parameter points to a directory where the config.json contains the correct `quantization\_config`. miles will automatically use FP8 quantization during weight updates.

3. Start FP8 training.

```
```bash
cd miles
Qwen3-4B FP8 training (single node)
bash examples/infra_features/low_precision/run-qwen3-4b-fp8.sh
Qwen3-30B-A3B FP8 training (two nodes)
bash examples/infra_features/low_precision/run-qwen3-30b-a3b-fp8-two-nodes.sh
```
```

Following the above command will launch FP8 training.

4. Use the saved checkpoint for evaluation.

Note that TransformerEngine does not specifically save FP8 quantized weights; the saved

torch dist remains in original precision (usually bf16). If you want to evaluate under FP8, you need to convert the checkpoint from `torch\_dist` to HuggingFace format, then convert to FP8 HuggingFace format.

--fp8-param-gather # [optional] Currently incompatible with CPU Adam `` Then ensure the NVTE_FP8_BLOCK_SCALING_FP32_SCALES` environment variable is enabled.

Note that only Linear and GroupLinear layers in TransformerEngine use fp8 format. embedding and lm_head remain in their original precision. If --fp8-param-gather is not enabled, weights in TransformerEngine remain in bf16, only being cast to fp8 during GEMM or GroupGEMM operations.

1. Convert your HuggingFace model weights to FP8 format.

You can use tools/convert_hf_to_fp8.py to convert bf16 weights to fp8 format. Ensure the --hf-checkpoint parameter points to a directory where the config.json contains the correct quantization_config. miles will automatically use FP8 quantization during weight updates.

1. Start FP8 training.

bash cd miles # Qwen3-4B FP8 training (single node) bash examples/infra_features/low_precision/run-qwen3-4b-fp8.sh # Qwen3-30B-A3B FP8 training (two nodes) bash examples/infra_features/low_precision/run-qwen3-30b-a3b-fp8-two-nodes.sh Following the above command will launch FP8 training.

1. Use the saved checkpoint for evaluation.

Note that TransformerEngine does not specifically save FP8 quantized weights; the saved torch dist remains in original precision (usually bf16). If you want to evaluate under FP8, you need to convert the checkpoint from torch_dist to HuggingFace format, then convert to FP8 HuggingFace format.

~~~~

## 评论区精华

PR 自身没有任何评论线程 (comments 与 review comments 均为 0) , 唯一审核来自 Zhichenzzz, 直接 APPROVED 且无正文, 说明这批文档修复没有引发争议。真正有信息量的内容沉淀在关联 PR #2481 的 body 中:

Stacked on #2482(examples README content fixes) — merge that first, then retarget this to main. The generator depends on those fixes: its dead-link check, required-H1 rule and description-length gate all fail against the unfixed READMEs.

这段文字点明了本 PR 作为镜像机制前置条件的地位: 不是可有可无的文档整理, 而是让 #2481 生成器的机器可检查规则 (死链、H1、描述长度) 能够通过的必要输入。另一个隐含决策来自 #2481 引述的 Slack 约定——examples/ 是唯一事实来源, 旧的手写文档页直接废弃; 本 PR 的内容修复正是这条约定的落地。

- 暂无高价值评论线程

## 风险与影响

- 风险：回归风险为零：纯文档变更，不触碰任何代码路径。主要风险有三类：一是隐性契约风险，examples/README.md 的描述文本在本 PR 之后会成为 #2481 生成器的 meta description 输入，未来若有人写了超长或格式不规则的条目描述，会触发 description-length 门控，这是一份“文档与自动化管道”之间的隐性契约；二是时效性风险，死链修复依赖人工审计，未来再有目录删除（如 #1953 这类）时链接仍可能再次失效，而 #2481 引入的 pre-commit hook 只覆盖 docs 站镜像的漂移检查，不兜底 GitHub 视图的链接可用性；三是合并顺序风险，若 #2481 先于本 PR 合入，其生成器的三条检查会失败。
- 影响：对 GitHub 读者：examples/ 顶层索引从“仓库布局”视角变为“读者导航”，全部示例（含实验性条目）可被发现；retool\_v2 读者不再被引向 404；low\_precision 的 Quick Start 步骤恢复可读。对团队：本 PR 是文档自动化管线的第一步，合并顺序敏感（必须先于 #2481）；它把“手工维护、必会漂移”的示例文档，变成 #2481 之后“自动生成 + drift 失败”的受控基建。影响范围仅限文档，不影响任何运行时行为。
- 风险标记：纯文档变更无代码风险，meta 描述契约，镜像门控前置，合并顺序敏感

## 关联脉络

- PR #2481 docs: mirror examples/ READMEs into the Examples tab: 直接栈上依赖：本 PR body 标注 "Stacked: #2481 (docs mirror mechanism) is based on this branch"; #2481 的 dead-link check、required-H1 rule 和 description-length gate 都以本 PR 修复后的 README 内容为前提，合并顺序必须先 #2482 后 #2481。
- PR #1953 Delete examples/retool: 本 PR 修复了 #1953 删除 examples/retool 后留下的死链引用（retool\_v2/README.md 仍指向该目录），是“删除目录后清理引用”的收尾工作。