

PR #2399 完整报告

radixark/miles

docs: fix the gsm8k download on the reproducibility page

合并时间: 2026-08-12 07:55

原文链接: <http://prhub.com.cn/radixark/miles/pull/2399>

执行摘要

- 一句话: 修复复现页 gsm8k 数据集下载命令
- 推荐动作: 变更小而明确, 不需要精读; 但其审计思路值得借鉴——核对文档声明的数据源与实际脚本读取的数据契约(路径 + 列名)。若你在处理同类「文档 / 脚本数据不一致」问题, 可参考本 PR 与 #2398、#2402 的联动处理方式。

功能与动机

PR body 明确指出这是对 #2398 的跟进, 来自同一轮对数据集下载命令的审计: reproducibility 页面与 Quick Start 页面存在同类 bug。文档页下载的 `openai/gsm8k` 文件位于 `main/train-00000-of-00001.parquet` (还有 `socratic/...`), 且只有 `question/answer` 列; 而页面随后运行的启动脚本 `examples/experimental/reproducibility/run-qwen2.5-0.5B-gsm8k.sh` 期望 `/root/gsm8k/train.parquet` 和 `/root/gsm8k/test.parquet`, 并以 `--input-key messages --label-key label` 读取列, 因此下载数据既不满足路径也不满足列名。`zhuzilin/gsm8k` 恰好提供根级 `train.parquet/test.parquet` 且列包含 `messages` 和 `label` (已通过 Hub tree 与 datasets server 验证), 并且 example 自己的 README 本来就让用户下载该数据集, 文档页是唯一不一致的地方。

实现拆解

1. 定位问题文件: 在 `docs/examples/reproducibility.md` 的「Download + convert + run」章节中, 发现下载命令指向 `openai/gsm8k`。
2. 核对数据契约: 对照 `examples/experimental/reproducibility/run-qwen2.5-0.5B-gsm8k.sh` 的参数 `--input-key messages --label-key label` 与文件路径期望, 确认 `openai/gsm8k` 不满足; 同时通过 Hugging Face Hub tree 与 datasets server 验证 `zhuzilin/gsm8k` 在根级提供 `train.parquet/test.parquet` 且列名匹配。
3. 单行修改: 将 `hf download --repo-type dataset openai/gsm8k --local-dir /root/gsm8k` 改为 `zhuzilin/gsm8k`, 与示例自带的 README 指引对齐。
4. 验证计划: PR body 列出两条手动验证命令(下载后确认 `train.parquet/test.parquet` 存在、运行启动脚本无路径或键错误), 未附自动化测试。
5. 评审与合入: 审核人 Zhichenzzz 直接 APPROVED, 无评论; 提交记录为单个 commit。

关键文件:

- docs/examples/reproducibility.md (模块 文档示例; 类别 docs; 类型 documentation) :
修复 gsm8k 数据集下载源: openai/gsm8k 的文件路径与列名和后续启动脚本期望不匹配, 改为 zhuzilin/gsm8k 后完全对齐。

关键符号: 未识别

关键源码片段

docs/examples/reproducibility.md

修复 gsm8k 数据集下载源: openai/gsm8k 的文件路径与列名和后续启动脚本期望不匹配, 改为 zhuzilin/gsm8k 后完全对齐。

```
# 修复前使用 openai/gsm8k, 其文件位于 main/train-00000-of-00001.parquet,  
# 且只有 question/answer 列, 与脚本期望的 train.parquet/test.parquet  
# 以及 messages/label 列不匹配, 导致复现流程失败。  
# 修复后改用 zhuzilin/gsm8k, 根级提供 train.parquet/test.parquet,  
# 列中包含 messages 和 label, 与脚本的 --input-key messages --label-key label 对应。  
hf download --repo-type dataset zhuzilin/gsm8k --local-dir /root/gsm8k
```

评论区精华

本 PR 无 review 评论或讨论线程, Zhichenzzz 直接 APPROVED。PR body 中的数据契约分析 (文件路径 + 列名必须与启动脚本匹配) 是本次变更的核心依据, 解释了为什么 openai/gsm8k 必然导致复现失败。

- 暂无高价值评论线程

风险与影响

- 风险: 主要风险来自对第三方社区数据集 zhuzilin/gsm8k 的依赖: 其目录结构和列名可能随上游变动而不再满足脚本契约; 不过它已被示例 README 长期引用, 且本次通过 Hub tree 与 datasets server 双重验证, 风险较低。除此之外, 变更涉及纯文档, 没有代码执行路径改动, 回归风险几乎为零。
- 影响: 影响范围限于 docs/examples/reproducibility.md 单页文档的一行命令。对希望按文档复现 gsm8k RL 训练的用户是直接修复, 消除了此前必然出现的文件路径与 --input-key/--label-key 键名错误; 对团队而言消除了文档页与 example README 之间的指引不一致。
- 风险标记: 第三方数据集依赖, 文档与脚本契约对齐

关联脉络

- PR #2402 scripts: download the DAPO dataset the NPU recipe trains on: 同类修复: 脚本 / 文档声明的数据集下载与训练脚本实际期望的数据路径错位, 体现同一轮数据契约对齐工作。