

PR #2354 完整报告

radixark/miles

Delete the glm4-9B, mimo-7B, moonlight-16B and deepseek-r1 launch scripts

合并时间: 2026-08-12 09:22

原文链接: <http://prhub.com.cn/radixark/miles/pull/2354>

执行摘要

- 一句话: 删除四个废弃模型启动脚本及配套快照与文档
- 推荐动作: 值得精读。本 PR 展示了如何在删除旧功能时进行精确的引用分析 (区分 " 真实引用 " 与 " 已禁用引用 ") , 以及如何处理文档与代码的联动清理。关键设计决策包括:
 - 通过 disabled 机制判断测试是否真实执行, 从而安全删除看似被引用的模型定义;
 - 保留仍被基础设施示例 (p2p、low-precision) 引用的模型定义, 避免过度删除;
 - 重写存活模型文档而非简单删除, 保证文档与实际 launcher 对齐。对维护大型脚本库的团队有很强的参考价值。

功能与动机

PR body 明确指出这些模型已被弃用 ("These four models are being dropped"), 因此需要移除对应的launcher、快照和文档。同时确保不误删仍被引用的模型定义 (如mimo-7B-rl.py、moonlight.py、deepseek-v3.py), 并通过 `pytest tests/manual/launch_scripts tests/fast/launch_scripts` 的 524 个测试证明无残留引用。

实现拆解

实现按以下步骤推进:

1. 删除废弃模型启动脚本与快照: 删除 `scripts/run-glm4-9B.sh`、`run-glm4-9B-4xgpu-radixtree.sh`、`run-mimo-7B-rl-eagle.sh`、`run-moonlight-16B-A3B.sh`、`run-deepseek-r1.sh` 及其对应的 `tests/snapshots/launch_scripts/sh/scripts/*.txt` 快照; 同时删除 `docs/models/glm/glm4.md`、`docs/models/mimo/mimo.md`、`docs/models/kiimi/moonlight.md` 三个文档页面和 `examples/infra_features/low_precision/run-moonlight-16B-A3B-int4.sh`。
2. 精确处理仍被引用的模型定义: 第一个提交保留 `scripts/models/mimo-7B-rl.py` (被 `tests/e2e/megatron/test_mimo_7B_mtp_only_grad.py` 注册使用)、`moonlight.py` (被 low-precision 示例和 p2p profile 引用)、`deepseek-v3.py` (被 `scripts/run_deepseek.py` 和 V3.2 launchers 引用)。第二个提交确认 `glm4-9B.py` 的唯一引用 `tests/e2e/megatron/test_quick_start_glm4_9B.py` 注册了 `disabled="Not worthy for testing, too naive."`, `run_suite.py` 只运行 `disabled` 为 `None` 的条目, 因此该测试从未执行, 故连同 p2p 中的 GLM-Z1-9B profile 一并删除。第三个提交确认 `moonlight.py`

的所有引用 (low-precision 脚本和 p2p profile) 都随本 PR 删除, 因此一并移除。

3. 重写 deepseek.md 并同步导航: DeepSeek-V3 仍受支持, 页面重写为围绕 `scripts/run_deepseek.py` 的真实参数 (TP/PP/CP/EP 阶梯、`deepseek low_latency` sglang 块、`--max-tokens-per-gpu 2048` 占位符), 删除 R1 专属内容, 更新 `docs/models/glm/index.md`、`docs/models/kimi/index.md`、`docs/models/index.md` 和 `docs/docs.json`。同时将 `docs/user-guide/training-script-walkthrough.md` 从引用已删除的 `run-glm4-9B.sh` 改指仍存活的脚本。
4. 修复快照漂移: 提交 `e5a9c267` 重新生成了 `run-qwen3-4B.sh` 的快照, 因为 #2300 添加了 `--use-miles-dashboard` 和 `--dump-details` 参数但未刷新快照, 而 `tests/manual` 未被 CI 发现, 导致差异长期存在于 main 上。
5. 合并与冲突解决: 两次合并 `origin/main` 解决 `docs/index.md`、`docs/models/deepseek/*`、`docs/models/glm/index.md` 等文档冲突; 最后一个提交恢复 DeepSeek V3 页面 (被 #2391 替换为 V3.2 配方但 V3 仍可训练), 并移除 #2391 的 redirect 以避免遮蔽恢复的页面。
6. 测试验证: 运行 `pytest tests/manual/launch_scripts tests/fast/launch_scripts` 共 524 个测试全部通过, 其中快照集合 == 脚本集合的 discovery 测试证明没有悬挂引用。

关键文件:

- `scripts/models/glm4-9B.py` (模块 模型定义; 类别 source; 类型 deletion; 符号 `model_args`): 模型定义文件, 确认其唯一引用测试被 disabled 后删除, 是引用分析的典型案例
- `tests/e2e/megatron/test_quick_start_glm4_9B.py` (模块 e2e 测试; 类别 test; 类型 deletion; 符号 `prepare, execute`): 该测试注册时带 disabled 标记, 证明从未执行, 是删除 glm4-9B 模型定义的关键依据
- `examples/infra_features/p2p_weight_transfer/run.py` (模块 p2p 示例; 类别 source; 类型 core-logic): 删除了 GLM-Z1-9B 和 Moonlight 的 `prepare/run` 配置, 是模型定义删除的连带变更
- `docs/models/deepseek/deepseek.md` (模块 模型文档; 类别 docs; 类型 documentation): 重写为围绕 `run_deepseek.py` 的文档, 是文档与代码对齐的典型修改, 且 review 讨论集中于此
- `scripts/run_deepseek.py` (模块 启动脚本; 类别 source; 类型 core-logic): 作为 `deepseek.md` 重写后的唯一 launcher, 其注释变更反映了对旧脚本行为的确认
- `scripts/run-deepseek-r1.sh` (模块 启动脚本; 类别 other; 类型 deletion): 被删除的五个启动脚本之一, 代表 R1 模型支持移除
- `scripts/run-mimo-7B-rl-eagle.sh` (模块 启动脚本; 类别 other; 类型 deletion): 被删除的五个启动脚本之一, 代表 MiMo-7B 支持移除
- `scripts/run-moonlight-16B-A3B.sh` (模块 启动脚本; 类别 other; 类型 deletion): 被删除的五个启动脚本之一, 代表 Moonlight-16B 支持移除
- `examples/infra_features/low_precision/run-moonlight-16B-A3B-int4.sh` (模块 低精度示例; 类别 other; 类型 deletion): 作为 Moonlight 的最后一个引用被删除, 导致 INT4 覆盖缺少廉价 smoke test

- `scripts/models/moonlight.py` (模块 模型定义; 类别 `source`; 类型 `deletion`; 符号 `model_args`) : 模型定义文件, 确认引用全部移除后删除

关键符号: `model_args`, `prepare`, `execute`

评论区精华

review 中只有两条评论, 集中在 [docs/models/deepseek/deepseek.md](https://docs.models.deepseek.com/deepseek.md):

- guapisolo 提出 [P1] 问题: 文档化命令在干净主机上会失败, 因为 `train()` 先调用 `_prepare_megatron_ckpt()` 和 `_prepare_cp()`, 两者最终调用 `exec_command_multi_node()`, 其第一步是 `ray.init(address="auto")`, 需要已存在的 Ray head; 而 `execute_train()` 默认会停止该集群 (除非设置 `MILES_SCRIPT_EXTERNAL_RAY=1`)。建议要么在预处理前建立并保留 Ray, 要么文档化外部 Ray 工作流。
- yueming-yuan 回复 "solved in next PR", 即该问题在后续 PR #2355/#2356 中解决, 状态为已接受续修。

整体 review 由 guapisolo APPROVED (LGTM), 无未解决的争议。

- DeepSeek V3 文档化命令的 Ray 生命周期问题 (correctness): 作者回复 "solved in next PR", 问题延期到后续 PR 修复。

风险与影响

- 风险: 主要风险在于删除引用链的完整性:
 - 如果存在未被发现的 `glm4-9B` / `moonlight` 模型类型引用, 将导致运行时找不到模型定义。PR 通过快照 discovery 测试 (524 个测试) 证明脚本集合与快照集合一致, 但 `e2e` 测试未运行, `run_suite.py` 的 `disabled` 过滤逻辑是本 PR 判断的核心依据, 若该过滤逻辑变化可能导致误删。
 - `examples/infra_features/p2p_weight_transfer/run.py` 删除了 `GLM-Z1-9B-0414` 和 `Moonlight-16B-A3B-Instruct` 的 `prepare/run` 配置, 若用户仍按旧文档使用该示例将报 `KeyError`; 文档同步更新降低了该风险。
 - `low_precision` 示例删除后, INT4 覆盖中最便宜的 smoke test (`Moonlight` 4 GPU 单节点) 丢失, 剩余最小 INT4 示例需要 8 GPU, 可能影响 CI 覆盖效率。
 - `docs/models/deepseek/deepseek.md` 重写后文档化命令存在 Ray 生命周期问题 (guapisolo P1 指出), 在下一个 PR 修复前, 按文档执行可能失败。
 - `run_deepseek.py` 的注释变更 (移除 `--use-kl-loss` 的 TODO 描述) 是行为相关的文案调整, 需确认无实际逻辑变化。
- 影响: 影响范围:
 - 对用户: 删除 5 个 shell 启动脚本和 3 个模型文档页, 用户如需训练 `GLM4-9B`、`MiMo-7B`、`Moonlight-16B` 或 `DeepSeek-R1` 将无官方入口; `DeepSeek-V3` 用户转向 `scripts/run_deepseek.py` (Python launcher) 和重写后的 `deepseek.md`。
 - 对系统: `examples/infra_features/p2p_weight_transfer/run.py` 的模型注册表缩小, `low_precision` 示例的 INT4 覆盖减少一个廉价用例。

- 对团队：这是三个串联清理 PR 的起点，减少维护面；快照重新生成修复了长期存在的文档 - 代码漂移，有助于后续自动发现此类问题。
- 风险标记：删除引用链依赖 disabled 过滤逻辑，文档命令存在 Ray 生命周期问题，INT4 覆盖丢失廉价 smoke test, 跨 PR 串联清理依赖

关联脉络

- PR #2391 docs: replace DeepSeek V3/R1 page with a DeepSeek-V3.2 recipe: 本 PR 的最后一个提交恢复了被 #2391 替换的 DeepSeek V3 页面，并移除了 #2391 的 redirect, 两者直接冲突并相互覆盖。
- PR #2300 scripts: enable the Miles dashboard in the quick-start launcher: 本 PR 重新生成 run-qwen3-4B.sh 的快照以纳入 #2300 添加的 dashboard 参数，修复了快照漂移。
- PR #2279 docs: polish the Quick Start page: 提交信息提到 tests/manual 未被 CI 发现，与 #2279 (CI 范围) 相关，导致快照漂移未被及时捕获。