PR #2716 完整报告

radixark/miles

fix: align MLA RoPE types with model configs

合并时间：2026-08-23 19:14

原文链接：http://prhub.com.cn/radixark/miles/pull/2716

执行摘要

- 一句话：为 GLM-4.7-Flash 与 JoyAI-LLM-Flash 显式声明 plain RoPE，修复 MLA 误用 YaRN
- 推荐动作：值得精读，尤其适合关注模型配置契约与训练框架隐式默认值交互的读者。本 PR 虽然只有 +12 行的最终改动，但完整呈现了一个隐蔽的配置错配 bug 的根因链条：模型定义遗漏参数 → 参数加载原样传递 → Megatron 侧隐式兜底 → 运行时构造错误的嵌入实现。值得借鉴的两个设计决策：一是"显式优于隐式"，在框架存在默认值填充逻辑时，模型定义应显式声明；二是合并前的 split-scope 纪律，未验证的 GLM-5 部分被果断剥离，避免把不确定性的风险带入主分支。

功能与动机

PR body 明确指出：python miles/utils/external_utils/model_args_utils.py glm4.7-flash 此前输出的参数只有 --multi-latent-attention 而没有 --rope-type，该遗漏被 load_model_args() 原样传播进 launcher 与转换命令；随后 set_default_megatron_args() 在 multi_latent_attention 启用时把缺失的 rope_type 映射为 yarn，最终让 MLA 实现构造出 YarnRotaryEmbedding 而非 RotaryEmbedding。这个兜底对 DeepSeek、Kimi 等确实使用 YaRN 的定义是合适的，但与 GLM-4.7-Flash、GLM-5/5.1/5.2 及 JoyAI-LLM-Flash 五个 HuggingFace 官方 checkpoint 描述的 plain/default RoPE 不符，属于静默的行为错配。

实现拆解

本 PR 的完整实现过程如下：

1. 定位根因链路：从 miles/utils/external_utils/model_args_utils.py 的复现命令入手，确认缺失 --rope-type 是问题起点；load_model_args() 原样传递该遗漏，set_default_megatron_args() 再以 yarn 兜底，最终 config.rope_type 驱动 MLA 位置编码构造分支。四个环节中只要模型定义显式声明 rope_type 即可切断错误传播。
2. 修改两个模型定义：在 scripts/models/glm4.7-flash.py 与 scripts/models/joyai-llm-flash.py 的 model_args() 返回串中，紧跟 --position-embedding-type rope 之后各插入一行 --rope-type rope。这是 data-contract 层面的修复：不触碰 Megatron 侧逻辑，也不改全局 MLA 默认值，仅让这两个模型的参数契约与其官方 checkpoint 对齐。

3. 同步刷新 7 个快照文件：覆盖三类快照——tests/snapshots/model_args/ 下的参数展开结果、tests/snapshots/launch_scripts/sh/examples/infra_features/p2p_weight_transfer/ 下的 shell launcher（GLM-4.7-Flash 2 节点 profile 场景）、以及 tests/snapshots/launch_scripts/py/scripts/run_glm47_flash.py/ 与 run_joy_ai_llm_flash.py/ 下 Python launcher 的 prepare/execute 各两组。快照更新验证了 --rope-type rope 能正确传播到 launcher 与 torchrun/ray job submit 命令的最终形态。
4. 收缩合并范围（第二个 commit）：合并者 guapisolo 将 GLM-5/5.1/5.2 的 plain-RoPE 变更从 PR 中移除，原因写在 commit message 与 issue 评论中：当前 GLM5 adapter 未验证 generic plain-RoPE override，直接切换会把 YarnRotaryEmbedding（forward() 返回两个值）换成 RotaryEmbedding（forward() 返回一个值），使 DSAMLASelfAttention.forward() 中无条件执行的 (rotary_pos_emb, mscale) 解包直接失败。GLM-4.7 与 JoyAI 的经验证部分保留。
5. 验证与配套：除快照外没有新增专门单元测试文件；合并者通过定向 GLM5 snapshot 检查（20 个 model-argument、28 个 Python-launcher、2 个 shell-launcher 用例）与 git diff --check 确认移除变更后仍保持一致性。这是以快照为回归护栏、以数据契约为修复对象的典型配套方式。

关键文件：

- scripts/models/glm4.7-flash.py（模块 模型定义；类别 source；类型 data-contract）：核心修复文件之一：在 model_args() 中紧跟 --position-embedding-type rope 后新增 --rope-type rope，切断 set_default_megatron_args() 对缺失 rope_type 的 yarn 隐式兜底。
- scripts/models/joyai-llm-flash.py（模块 模型定义；类别 source；类型 data-contract）：核心修复文件之二：同样的 data-contract 修复，为 JoyAI-LLM-Flash 的 MLA 配置显式声明 --rope-type rope，避免其误用 YaRN。
- tests/snapshots/model_args/glm4.7-flash.txt（模块 快照测试；类别 docs；类型 documentation）：关键回归快照：验证 model_args() 的输出中正确出现 --rope-type rope 键值对，是参数契约修复的直接证据。
- tests/snapshots/model_args/joyai-llm-flash.txt（模块 快照测试；类别 docs；类型 documentation）：JoyAI-LLM-Flash 的参数展开快照，验证 --rope-type rope 进入最终参数集。
- tests/snapshots/launch_scripts/py/scripts/run_glm47_flash.py/prepare.txt（模块 快照测试；类别 docs；类型 documentation）：验证 torchrun prepare 阶段命令已携带新参数，说明修复能完整传播到实际训练入口。
- tests/snapshots/launch_scripts/py/scripts/run_glm47_flash.py/execute.txt（模块 快照测试；类别 docs；类型 documentation）：验证 ray job submit 执行阶段命令同样携带新参数，覆盖 launch 全链路。
- tests/snapshots/launch_scripts/py/scripts/run_joy_ai_llm_flash.py/prepare.txt（模块 快照测试；类别 docs；类型 documentation）：JoyAI-LLM-Flash 的 prepare 阶段命令快照，验证 torchrun 入口携带新参数。
- tests/snapshots/launch_scripts/py/scripts/run_joy_ai_llm_flash.py/execute.txt（模块 快照测试；类别 docs；类型 documentation）：JoyAI-LLM-Flash 的 ray job submit 执行阶

段命令快照。

- tests/snapshots/launch_scripts/sh/examples/infra_features/p2p_weight_transfer/run-glm4.7-flash-2node-profile.sh.txt（模块 快照测试；类别 docs；类型 documentation）：p2p_weight_transfer 场景的 shell launcher 快照，覆盖 2 节点 profile 用例，验证 sh 入口同样携带新参数。

关键符号：model_args (scripts/models/glm4.7-flash.py), model_args (scripts/models/joyai-llm-flash.py)

关键源码片段

scripts/models/glm4.7-flash.py

核心修复文件之一：在 model_args() 中紧跟 --position-embedding-type rope 后新增 --rope-type rope，切断 set_default_megatron_args() 对缺失 rope_type 的 yarn 隐式兜底。

```
# scripts/models/glm4.7-flash.py 中的 model_args()，修复后版本
# 修复要点：此前只声明了 --position-embedding-type rope 却没有 --rope-type，
# 一旦启用 --multi-latent-attention，set_default_megatron_args() 会把缺失的
# rope_type 隐式填充为 yarn，与官方 checkpoint 的 plain/default RoPE 不一致。
def model_args() -> str:
    return (
        f"--moe-layer-freq [0]*{N_DENSE_LAYERS}+[1]*{N_MOE_LAYERS} "
        f"--num-experts {MOE_ROUTED_EXPERTS} "
        f"--moe-shared-expert-intermediate-size {MOE_SHARED_EXPERT_INTERMEDIATE_SIZE} "
        f"--moe-router-topk {MOE_ACTIVE_ROUTED_EXPERTS} "
        "--moe-grouped-gemm "
        "--moe-permute-fusion "
        f"--moe-ffn-hidden-size {MOE_FFN_HIDDEN} "
        "--moe-router-score-function sigmoid "
        "--moe-router-pre-softmax "
        "--moe-router-enable-expert-bias "
        "--moe-router-bias-update-rate 0 "
        "--moe-router-load-balancing-type seq_aux_loss "
        "--moe-router-topk-scaling-factor 1.8 "
        "--moe-aux-loss-coeff 0 "
        "--moe-router-dtype fp32 "
        "--make-vocab-size-divisible-by 64 "
        f"--num-layers {N_DENSE_LAYERS + N_MOE_LAYERS} "
        f"--hidden-size {NHIDDEN} "
        f"--ffn-hidden-size {FFN_HIDDEN} "
        f"--num-attention-heads {NHEADS} "
        "--disable-bias-linear "
        "--add-qkv-bias "
        "--swiglu "
        "--untie-embeddings-and-output-weights "
        "--position-embedding-type rope "
        "--rope-type rope " # 显式声明 plain RoPE，阻断 yarn 隐式默认值
        "--no-position-embedding "
```

```
        "--normalization RMSNorm "
        "--norm-epsilon 1e-5 "
        "--qk-layernorm "
        "--multi-latent-attention "
        "--q-lora-rank 768 "
        "--kv-lora-rank 512 "
        "--qk-head-dim 192 "
        "--v-head-dim 256 "
        "--kv-channels 192 "
        "--qk-pos-emb-head-dim 64 "
        "--vocab-size 154880 "
        "--rotary-base 1000000 "
        "--no-rope-fusion "
        "--mtp-num-layers 1 "
    )
```

scripts/models/joyai-llm-flash.py

核心修复文件之二：同样的 data-contract 修复，为 JoyAI-LLM-Flash 的 MLA 配置显式声明 `--rope-type rope`，避免其误用 YaRN。

```
# scripts/models/joyai-llm-flash.py 中的 model_args()，修复后版本
# 与 GLM-4.7-Flash 相同：在 MLA 参数组之前显式声明 --rope-type rope，
# 否则该模型会从 set_default_megatron_args() 继承 yarn 兜底值。
def model_args(nlayers: int | None = None, rotary_base: str | None = None) -> str:
    nlayers = nlayers if nlayers is not None else int(os.environ.get("MODEL_ARGS_NUM_
    LAYERS") or 40)
    rotary_base = rotary_base if rotary_base is not None else os.environ.get("MODEL_ARGS_
    ROTARY_BASE") or "32000000"
    return (
        "--disable-bias-linear "
        f"--num-layers {nlayers} "
        "--hidden-size 2048 "
        "--ffn-hidden-size 7168 "
        "--num-attention-heads 32 "
        "--kv-channels 128 "
        "--normalization RMSNorm "
        "--position-embedding-type rope "
        "--rope-type rope " # 显式声明 plain RoPE，与官方 checkpoint 对齐
        "--norm-epsilon 1e-6 "
        "--swiglu "
        "--untie-embeddings-and-output-weights "
        "--vocab-size 129280 "
        "--multi-latent-attention "
        "--q-lora-rank 1536 "
        "--kv-lora-rank 512 "
        "--qk-head-dim 128 "
        "--qk-pos-emb-head-dim 64 "
        "--v-head-dim 128 "
        "--qk-layernorm "
```

```
        f"--rotary-base {rotary_base} "
        "--mscale 1.0 "
        "--mscale-all-dim 1.0 "
        "--attention-softmax-in-fp32 "
        "--no-rope-fusion "
        "--num-experts 256 "
        f"--moe-layer-freq {moe_layer_freq(nlayers=nlayers, first_k_dense_replace=FIRST_K_
        DENSE_REPLACE)} "
        "--moe-ffn-hidden-size 768 "
        "--moe-router-topk 8 "
        "--moe-shared-expert-intermediate-size 768 "
        "--moe-router-pre-softmax "
        "--moe-router-score-function sigmoid "
        "--moe-router-enable-expert-bias "
        "--moe-router-load-balancing-type seq_aux_loss "
        "--moe-token-dispatcher-type alltoall "
        "--moe-aux-loss-coeff 0 "
        "--moe-router-bias-update-rate 0 "
        "--moe-router-group-topk 1 "
        "--moe-router-num-groups 1 "
        "--moe-grouped-gemm "
        # 其余 MoE 路由参数与原始定义一致，此处省略
    )
```

评论区精华

本 PR 没有 inline review 评论，核心讨论集中在 Issue 评论区与 PR body 的 Note 段落：

- 合并者 guapisolo 在 Issue 评论中提到“Letting my agent to do ablation and see the scale of rope issue”，先用自动化手段评估 rope 问题的影响规模，再决定变更范围。
- 在第二个 commit 的说明中，guapisolo 明确解释了收缩范围的理由：

Commit b9f61310 will remove the GLM5/5.1/5.2 plain-RoPE launcher and generated snapshot changes from this PR because the current GLM5 adapter has not been validated for that override, while preserving the GLM-4.7 and JoyAI changes.

- PR body 的 Note 是本次最重要的技术论证：

GLM-5 DSAMLASelfAttention.forward() unconditionally unpacks (rotary_pos_emb, mscale), matching the two-value return contract of YarnRotaryEmbedding.forward(). Selecting --rope-type rope instead constructs RotaryEmbedding, whose forward() returns one tensor, so the existing GLM-5 path can fail during unpacking.

- 结论：GLM-5 的适配必须与 attention 调用点一起改造后才能合并，本次先合并已验证的 GLM-4.7-Flash 与 JoyAI-LLM-Flash，GLM-5 作为 follow-up 保留。

- GLM-5 系列是否纳入本次 plain-RoPE 修复 (design): GLM-5 变更被移除，仅保留已验证的 GLM-4.7 与 JoyAI 部分；GLM-5 需要 attention 调用点一起适配后才能安全合并。
- 缺失 rope_type 时 yarn 兜底的契约风险 (correctness): 采用数据契约层面的修复：仅给受影响模型加 --rope-type rope，不改全局 MLA 默认值，不影响显式 YaRN 配置。

风险与影响

- 风险：
  1. 位置编码语义切换风险（核心）：scripts/models/glm4.7-flash.py 与 scripts/models/joyai-llm-flash.py 切换后，Megatron 侧将从 YarnRotaryEmbedding 变为 RotaryEmbedding。PR 未改动这两个模型对应的 attention adapter，如果它们也像 GLM-5 那样对 rotary 输出做双值解包，会在运行时报错；不过合并前 GLM-4.7 与 JoyAI 的快照与定向校验已通过，风险相对可控。
1. 与既有训练产物兼容性：若此前用隐式 yarn 配置启动过训练或生成过 checkpoint，切换为 plain RoPE 后位置嵌入初始化不同，续训或加载旧权重会出现位置编码语义漂移，需要重新对齐或重新初始化。
2. 测试覆盖缺口：没有新增直接针对 rope_type 传播的单元测试文件，回归保护完全依赖快照。快照覆盖了 model_args、shell launcher、Python launcher 的 prepare/execute 全链路，但仍缺少对 set_default_megatron_args() 兜底逻辑本身的单元断言。
3. 派生变体继承：pruned 与 LoRA 变体自动继承修正值，但若某些派生脚本存在自己的参数覆盖逻辑，存在被二次覆盖的隐患；本次快照未覆盖所有派生组合。 - 影响：对用户的影响：运行 GLM-4.7-Flash 与 JoyAI-LLM-Flash 的训练、转换流程时，生成的 launcher 命令会多出 --rope-type rope，位置编码从隐式 YaRN 切换为 plain RoPE，更贴近 HuggingFace 官方 checkpoint 的语义。对系统的影响：全局 MLA 默认值与显式 YaRN 配置（DeepSeek、Kimi）完全不受影响，改动是模型级、局部化的数据契约修复。对团队的影响：确立了“MLA 模型定义必须显式声明 rope_type”的约定，并示范了在未验证的模型变体面前收缩合并范围的协作方式；GLM-5 系列留下明确的 follow-up 任务。 - 风险标记：隐式默认值兜底链路，位置编码语义切换，GLM-5 适配未完成，缺少新增单元测试

关联脉络

- PR #2682 fix: resume from the checkpoint step in bridge mode: 同处模型参数处理链路：2682 修复 Megatron 参数在桥接模式下的检查点恢复问题，本 PR 修复模型参数定义中 rope_type 的契约缺失，二者都属于训练参数正确性主题。
- PR #2588 glm52_tbench2: collapse levers and launch fixes from the GB300 16-node smoke runs: 同为 GLM 系列训练稳定性修复：2588 处理 GLM5.2 的启动参数与注意力前端问题，与本 PR 中暂缓的 GLM-5 适配直接相关，构成 GLM5 adapter 演进的一条连续线索。