PR #2280 完整报告

radixark/miles

[example] GLM-5.2 744B-A40B LoRA agentic launcher (TB2 on Daytona)

合并时间：2026-08-13 05:38

原文链接：http://prhub.com.cn/radixark/miles/pull/2280

执行摘要

- 一句话：新增 GLM-5.2 744B LoRA agentic 多节点启动器示例
- 推荐动作：值得精读，尤其是对多节点 LoRA + agentic 训练且基础模型超过单机显存的用户。脚本中的 DSA 后端选择（tilelang vs megatron）、磁盘 offload 与 allocator env 设置均附有详细注释，设计决策清晰，可作为编写其他大型模型启动器的范本。

功能与动机

PR 正文说明：Neither existing example covers this combination: run.py / run-glm47-flash-agentic-async.py handle the agentic Harbor path for a model that fits on one node, and scripts/run_glm5_2_744b_a40b_lora.py handles GLM-5.2 LoRA on non-agentic data. The 744B base does not fit on a single 8x H200 node in bf16, so the recipe that makes the agentic + LoRA combination work is non-obvious and worth checking in as an example. 因此需要把这一组合的可行配置沉淀为可复用示例。

实现拆解

1. 新增 examples/swe-agent-harbor-docker/run_glm52_lora_tb2_daytona.py：定义 ScriptArgs 配置类，集中管理训练（lora_rank、target_modules、global_batch_size 等）、rollout（sglang_context_length、fp8_rollout_gpus_per_engine 等）、agent（agent_server_url、harbor_tasks_dir 等）与观测（W&B、Prometheus）参数；默认值已针对 4 节点 × 8 H200 调好。
2. 实现 _parallel_args：生成 Megatron 并行参数，TP 固定为节点内 GPU 数（NVLink），EP 覆盖全 world，PP=CP=1；根据 DSA 后端选择 qkv 格式（tilelang 用 thd，megatron 用 bshd），并因 bshd 不支持 dynamic-batch-size 而强制 --micro-batch-size 1。
3. 实现 _sglang_args 与 _write_sglang_fp8_config：生成 rollout 侧 sglang 服务配置，加载 fp8 checkpoint，开启 update_weights: true 使每步 LoRA 权重同步到达引擎；同时设置 --sglang-lora-backend triton（csgmv 在 DSA MoE-LoRA + dp-attention 下会崩溃）。
4. 配置训练器 offload：--offload-train-target disk 且目录指向非 tmpfs 的 /scratch/miles_train_offload，原因是暂停的 actor 每 rank 约 50 GiB，若做 pinned 主机副本会在 2 TiB 节点上 OOM；磁盘卸载将主机占用压到每 rank 仅一个 chunk。
5. 提供 cleanup 与 execute：清理陈旧 sglang / train 进程，组装完整命令行并执行；支持 debug_rollout_only 模式便于 smoke 验证。

6. 更新 README.md 文件表，登记该启动器一行说明。

关键文件：

- examples/swe-agent-harbor-docker/run_glm52_lora_tb2_daytona.py（模块 示例启动器；类别 infra；类型 infrastructure；符号 ScriptArgs, cleanup, _parallel_args, _sglang_args）：新增的多节点 GLM-5.2 744B-A40B LoRA agentic 启动器，包含全部关键并行配置、offload 策略和 agent 接入逻辑，是本 PR 的核心交付。
- examples/swe-agent-harbor-docker/README.md（模块 示例文档；类别 docs；类型 documentation）：在示例目录文件表中登记新启动器，一行说明其定位，便于用户发现。

关键符号：ScriptArgs, cleanup, _parallel_args, _sglang_args, _write_sglang_fp8_config, execute, main

评论区精华

PR 无 review 评论，yushengsu-thu 直接批准。作者在 PR 描述中自行说明了几个关键决策：

- 移除 MILES_TRAIN_MEMORY_FRACTION 环境变量，因为其指向的训练器侧补丁 #2199 未合并，变量实际无人读取；同时警告不要在 job-wide ray runtime env 里加 per_process_memory_fraction，否则 colocated rollout 引擎会继承此上限并 OOM 在 --sglang-mem-fraction-static 之下。
- 暂无高价值评论线程

风险与影响

- 风险：
  1. 示例脚本依赖 tilelang sparse-MLA 的 padded-index guard（#2079），若环境未包含该修复将出现错误；
  2. Test plan 中 --mode debug_rollout_only 的 smoke 尚未执行，脚本未在主分支构建上验证；
  3. 配置高度依赖特定硬件（8×H200、2 TiB 节点、非 tmpfs 目录），换环境需手工调整；
  4. 脚本仅作为示例提供，不参与 CI，配置漂移风险由使用者自行承担。 - 影响：影响范围限于 examples/swe-agent-harbor-docker/ 目录，对核心训练 /rollout 代码无任何改动。为需要多节点、agentic LoRA 训练的用户提供了可直接参考的完整配方，尤其解决了 744B 等超大模型无法单节点承载时的并行与内存管理问题。团队可借助该示例快速复现相近规模的 agentic RL 实验。 - 风险标记：依赖未合并的 #2079, debug_rollout_only smoke 未完成，环境配置敏感（硬件 / 路径 / 非 tmpfs），无测试覆盖

关联脉络

- PR #2079 tilelang sparse-MLA padded-index guard: 本启动器的 tilelang DSA 路径依赖该 PR 提供的 padded-index 保护，作者在 Notes for reviewers 中明确指出。
- PR #2199 trainer-side allocator cap companion patch: 该 PR 未合并，导致启动器移除了 MILES_TRAIN_MEMORY_FRACTION 环境变量，并改为在注释中建议用 torch.cuda.set_per_process_memory_fraction 兜底。

- PR #2368 fix(rollout): group session v2 leaf samples: 修复了 agentic session v2 的叶子样本分组与失败保护，本启动器使用的 agentic_tool_call.generate 依赖该修复。