

PR #7594 完整报告

verl-project/verl

[recipe, cfg] fix: enable sleep mode for KV cache release

合并时间: 2026-08-31 10:27

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7594>

执行摘要

本 PR 修复了 `verl/experimental/fully_async_policy/` 目录下几个启动脚本中错误的 vLLM 配置，将 `enable_sleep_mode` 从 `False` 更正为 `True` 或移除显式禁用设置。这是确保基于 PR#7373 引入的 KV 缓存回收（使用 `sleep/wake_up` API）功能正常运行的关键修复，防止 rollout 因配置不兼容而失败。

功能与动机

根据 PR body 阐述，此变更是对 PR#7373 的后续修复。PR#7373 实现了 vLLM 通过 `sleep` 和选择性唤醒来释放 KV 缓存。然而，当 vLLM 配置中的 `enable_sleep_mode` 被禁用时，它会拒绝相关的 `sleep/wake_up` API 调用，这将导致 rollout 流程失败。本 PR 的动机就是纠正之前脚本中遗留的不兼容配置，以保证 KV 缓存回收功能在 `fully-async` 场景下可用。

实现拆解

变更非常集中且直接，全部是配置文件调整：

- 配置修正 (`dapo_30b_a3b_base_math_fsdp.sh`): 在指定 vLLM 参数的部分，将 `enable_sleep_mode=False` 修改为 `enable_sleep_mode=True`。这是明确启用必要功能。
- 移除错误配置 (`geo3k_qwen3vl_30b_a3b_fsdp2_32_32.sh`, `dapo_30b_a3b_math_fsdp_npu.sh`): 这两个脚本显式设置了 `actor_rollout_ref.rollout.enable_sleep_mode=False`，此行被完全移除。由于该参数的默认值为 `True`，移除后即恢复了正确状态。
- 文件格式修复 (`geo3k_qwen3vl_30b_a3b_fsdp2_32_32.sh`): 同时修复了该文件末尾缺失的换行符，确保格式规范。

所有变更均位于 `verl/experimental/fully_async_policy/shell/` 下的实验性启动脚本，不影响核心代码库。

评论区精华

无。PR 审核过程中未产生讨论，由 wuxibin89 直接批准合并。

风险与影响

- 风险：主要风险在于“配置遗漏”。本次修复仅针对提交的三个具体脚本。如果用户拥有其他未更新的自定义启动脚本，且其中同样配置了 `enable_sleep_mode=False`，rollout 将仍然失败。这要求清晰的文档更新和社区通知，以告知所有使用 KV 缓存回收功能的用户必须启用

sleep mode。

- 影响：对于已经部署或计划使用这三个特定 fully-async 脚本的用户，此 PR 解除了功能阻断问题。影响范围有限（实验性功能），但影响程度高（功能性失败）。

关联脉络

此 PR 是 PR#7373([vllm, rollout] feat: add vLLM consumer for delta-sharded weight sync) 的直接后续和修复。PR#7373 引入了核心的稀疏权重同步和 KV 缓存管理机制，本 PR 确保了配套脚本能正确配置以利用该机制。它属于 vLLM 集成与性能优化功能线的一部分，后续可能需要更广泛的文档或脚本模板更新来强化这一配置要求。