PR #7558 完整报告

verl-project/verl

[ci] fix: drop stale enable_chunked_prefill=False from Ascend NPU scripts

合并时间：2026-08-26 15:58

原文链接：http://prhub.com.cn/verl-project/verl/pull/7558

执行摘要

- 一句话：移除 Ascend 脚本中失效的 chunked prefill 配置
- 推荐动作：值得快速合并，属于必要的清理修复。无需深入阅读源码，但可关注 PR #7508 的关联性。

功能与动机

PR #7508 修复了 CLI 序列化丢弃显式 False 布尔参数的问题，导致原本被静默忽略的 enable_chunked_prefill=False 开始生效，进而使 vLLM 启动中止。本 PR 旨在移除这些过时配置，恢复 Ascend NPU 脚本的正常运行，详见 PR body 中引用的 CI 运行记录。

实现拆解

1. 定位并修改四个 Ascend NPU 脚本，删除其中的 actor_rollout_ref.rollout.enable_chunked_prefill=False 行。
2. 涉及文件为 tests/special_npu/nightly_ci_ascend/run_dapo_moonlight-16b_megatron_npu.sh、tests/special_npu/quick_start/run_qwen3_0_6b_fsdp2_vllm_ascend.sh、tests/special_npu/quick_start/run_qwen3_0_6b_megatron_vllm_ascend.sh 和 tests/special_npu/run_qwen3_vl_8b_Instruct_fsdp2_npu.sh。
3. 每处仅删除一行，无其他改动，属于纯配置清理。

关键文件：

- tests/special_npu/nightly_ci_ascend/run_dapo_moonlight-16b_megatron_npu.sh（模块 Ascend 脚本；类别 test；类型 test-coverage）：删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止
- tests/special_npu/quick_start/run_qwen3_0_6b_fsdp2_vllm_ascend.sh（模块 Ascend 脚本；类别 test；类型 test-coverage）：删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止
- tests/special_npu/quick_start/run_qwen3_0_6b_megatron_vllm_ascend.sh（模块 Ascend 脚本；类别 test；类型 test-coverage）：删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止
- tests/special_npu/run_qwen3_vl_8b_Instruct_fsdp2_npu.sh（模块 Ascend 脚本；类别 test；类型 test-coverage）：删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止

关键符号：未识别

关键源码片段

tests/special_npu/nightly_ci_ascend/run_dapo_moonlight-16b_megatron_npu.sh

删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止

```
#!/bin/bash
# 之前包含 enable_chunked_prefill=False，现已移除
# 此配置在 PR #7508 修复后会导致 vLLM 启动失败

actor_rollout_ref.rollout.gpu_memory_utilization=0.65 \
actor_rollout_ref.rollout.tensor_model_parallel_size=${INFER_TP} \
actor_rollout_ref.rollout.data_parallel_size=1 \
# enable_chunked_prefill=False 已删除
actor_rollout_ref.rollout.max_num_batched_tokens=$((max_prompt_length + max_response_
length)) \
```

tests/special_npu/quick_start/run_qwen3_0_6b_fsdp2_vllm_ascend.sh

删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止

```
#!/bin/bash
# 从 ROLLOUT 配置中移除 enable_chunked_prefill=False

ROLLOUT=(
    actor_rollout_ref.rollout.name=vllm
    actor_rollout_ref.rollout.tensor_model_parallel_size=${ROLLOUT_TP}
    actor_rollout_ref.rollout.gpu_memory_utilization=${rollout_gpu_mem_util}
    # enable_chunked_prefill=False 已删除
    actor_rollout_ref.rollout.n=${ROLLOUT_N}
)
```

tests/special_npu/quick_start/run_qwen3_0_6b_megatron_vllm_ascend.sh

删除过时的 enable_chunked_prefill=False，修复 vLLM 启动中止

```
#!/bin/bash
# 从 ROLLOUT 配置中移除 enable_chunked_prefill=False

ROLLOUT=(
    actor_rollout_ref.rollout.log_prob_micro_batch_size_per_gpu=1
    # enable_chunked_prefill=False 已删除
    actor_rollout_ref.rollout.tensor_model_parallel_size=2
)
```

评论区精华

无 review 讨论，仅 wucong25 的 APPROVED 审核，未提出异议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，因为移除的是原本无效的配置，且在 PR #7508 修复后才暴露问题。但需注意，若某些脚本依赖该配置（尽管无效），移除后行为可能变化，不过从代码看默认值应为 False，因此风险可控。需在实际 Ascend 环境验证脚本能正常启动。
- 影响：影响范围仅限于 Ascend NPU 测试脚本，修复 vLLM 启动崩溃问题，对生产代码无影响。对维护者而言，减少了死配置，提高脚本可维护性。
- 风险标记：配置清理 , 依赖上游 bugfix

关联脉络

- PR #7508 [vllm] fix: honor explicit False on Optional[bool] engine args in CLI serialization: 该 PR 修复了 CLI 序列化问题，导致本 PR 中的 enable_chunked_prefill=False 开始生效并引发问题，从而需要清理。