PR #7632 完整报告

verl-project/verl

[vllm] fix: raise max_num_batched_tokens to max_model_len when chunked prefill is disabled

合并时间：2026-08-31 18:43

原文链接：http://prhub.com.cn/verl-project/verl/pull/7632

执行摘要

- 一句话：修复禁用 chunked prefill 时 vLLM 启动崩溃并清理旧配置
- 推荐动作：值得精读。核心价值不在 diff 规模，而在 _validate_configs 中如何用“warning + 自动提升”平衡 vLLM 约束与用户显式配置；同时展示了跨示例、CI、测试的连锁清理方法，以及测试在行为不确定性下如何改为观察式断言。建议重点关注 vllm_async_server.py 的安全网分支和 agent-loop 测试的改写思路。

功能与动机

PR body 明确指出：vLLM 在 enable_chunked_prefill=False 时要求 max_num_batched_tokens >= max_model_len，而 verl 始终向 vLLM 显式传入 max_num_batched_tokens（默认 8192），覆盖了 vLLM 自身在非 chunked 场景下默认取 max_model_len 的行为，导致 Qwen2.5-VL-3B-Instruct 等长上下文模型启动时抛出 ValidationError: max_num_batched_tokens (8192) is smaller than max_model_len (128000)。该问题由 geo3k VLM E2E 测试在 vllm 0.24 上触发。同时，示例脚本中残留的 enable_chunked_prefill=False 原本是绕开 vLLM VLM 占位符缺陷（vllm#15185，2025 年 3 月已修复）的 workaround，继续保留会对新版本造成启动失败，需随 #7629/#7624 一并清理。

实现拆解

1. 引擎侧安全网（核心修复）：在 verl/workers/rollout/vllm_rollout/vllm_async_server.py 的 _validate_configs 中新增分支：当 enable_chunked_prefill=False 且 max_num_batched_tokens < max_model_len 时，先通过 logger.warning 说明原因，再将 max_num_batched_tokens 就地提升为 max_model_len。该逻辑对齐 vLLM 自身的默认行为，且只影响非 chunked 场景；用户显式配置的更大值不会被覆盖。sglang / trtllm 后端不传递该参数，不受影响。
2. 示例脚本清理：沿 #7629/#7624 的思路，从 examples/grpo_trainer 下的 run_glm4_1v_9b_fsdp.sh、run_minicpm_o_2_6_fsdp.sh、run_qwen2_5_32b_fsdp.sh、run_qwen2_5_vl_7b_fsdp.sh、run_qwen3_4b_fsdp.sh、run_qwen3_5_2b_openr1_fsdp.sh、run_qwen3_vl_30b_moe_veomni.sh、run_qwen3_vl_8b_fsdp.sh 以及 examples/profile 下的 run_qwen2_5_vl_7b_torch_memory.sh、run_qwen3_8b_npu_profile_discrete.sh 中删除 enable_chunked_prefill=False；由于 rollout.yaml 已默认 True，示例恢复开箱即用。

torch profiler 脚本 run_qwen2_5_7b_torch_profile.sh 为获取干净 trace 保留禁用，并补上 max_model_len 显式推导。NPU 脚本均未改动。

3. E2E 与 CI 配置清理：tests/special_e2e/ppo_trainer/run_function_reward.sh 移除对 vllm#15185 的注释引用，geo3k 路由不再强制 ENABLE_CHUNKED_PREFILL=False；tests/special_e2e/run_ppo_trainer_veomni.sh、tests/special_e2e/run_ppo_trainer_torchtitan.sh 同步删除参数；.github/workflows/e2e_ppo_trainer_megatron_sglang_2.yml 与 e2e_ppo_trainer_megatron_vllm_2.yml 中的 geo3k 任务也移除了该环境变量。
4. agent-loop 负载均衡测试适配：#7613 将并列最少负载的副本选择改为随机，tests/experimental/agent_loop/test_basic_agent_loop.py 中 test_new_requests_route_to_least_loaded、test_get_inflight_count、test_removed_server_invalidates_sticky_session 原来硬编码 s0/s1/s2，现改为先观察实际分配副本再断言，消除 flaky。
5. 配套格式修复：verl/utils/model.py 在 from transformers import PreTrainedModel 后补上缺失空行，满足 ruff format 的 CI 检查。

关键文件：

- verl/workers/rollout/vllm_rollout/vllm_async_server.py（模块 引擎校验；类别 source；类型 core-logic；符号 _validate_configs）：核心修复所在：在引擎配置校验阶段自动提升 max_num_batched_tokens，解决非 chunked 模式下长上下文模型的启动崩溃，是本次变更的主路径。
- tests/experimental/agent_loop/test_basic_agent_loop.py（模块 负载均衡；类别 test；类型 test-coverage；符号 test_new_requests_route_to_least_loaded, test_removed_server_invalidates_sticky_session, test_get_inflight_count）：适配 #7613 随机平局打破，修复三个 GPU CI 测试的 flaky 问题，是本次除引擎修复外的第二处关键改动。
- verl/utils/model.py（模块 工具类；类别 source；类型 other）：修复 #7625 引入的格式问题，补上缺失空行以通过 ruff format CI，属于配套性改动。
- tests/special_e2e/ppo_trainer/run_function_reward.sh（模块 E2E 脚本；类别 test；类型 configuration）：geo3k VLM 路由的入口脚本：移除 vllm#15185 注释引用与强制 ENABLE_CHUNKED_PREFILL=False，让 E2E 回归默认配置。
- tests/special_e2e/run_ppo_trainer_veomni.sh（模块 E2E 脚本；类别 test；类型 configuration）：veomni 后端 E2E 脚本删除 enable_chunked_prefill=False 覆盖，随默认值走 chunked prefill。
- tests/special_e2e/run_ppo_trainer_torchtitan.sh（模块 E2E 脚本；类别 test；类型 configuration）：torchtitan 后端 E2E 脚本同步删除 enable_chunked_prefill=False，保持示例配置统一。
- examples/profile/run_qwen2_5_7b_torch_profile.sh（模块 分析脚本；类别 other；类型 configuration）：profiler 脚本有意保留禁用 chunked prefill 以获取干净 trace，同时补上 max_model_len 显式推导，避免触发启动校验崩溃。
- .github/workflows/e2e_ppo_trainer_megatron_vllm_2.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CI 中 geo3k VLM E2E 路由移除

ENABLE_CHUNKED_PREFILL=False，与新默认值对齐。

关键符号：_validate_configs, test_new_requests_route_to_least_loaded, test_removed_server_invalidates_sticky_session, test_get_inflight_count

关键源码片段

verl/workers/rollout/vllm_rollout/vllm_async_server.py

核心修复所在：在引擎配置校验阶段自动提升 max_num_batched_tokens，解决非 chunked 模式下长上下文模型的启动崩溃，是本次变更的主路径。

```
def _validate_configs(self) -> None:
    """校验 config / model_config，并修正与 vLLM 启动约束冲突的参数。"""
    max_position_embeddings = get_max_position_embeddings(self.model_config.hf_config)
    if self.config.max_model_len is None:
        # 未显式指定时，回退到模型位置编码的最大长度
        self.config.max_model_len = max_position_embeddings
    else:
        if self.config.max_model_len > max_position_embeddings:
            raise ValueError(
                f"max_model_len ({self.config.max_model_len}) should be less than or equal to "
                f"max_position_embeddings ({max_position_embeddings})"
            )

    # 关闭 chunked prefill 时，vLLM 要求单个 prefill 必须能放进一次调度迭代，
    # 因此必须满足 max_num_batched_tokens >= max_model_len。
    # verl 总是显式传入 max_num_batched_tokens（默认 8192），
    # 反而覆盖了 vLLM 自身在非 chunked 场景下“默认为 max_model_len”的行为，
    # 导致 128k 长上下文模型在引擎启动时直接抛错。
    if not self.config.enable_chunked_prefill and self.config.max_num_batched_tokens < self.
    config.max_model_len:
        logger.warning(
            "enable_chunked_prefill=False requires max_num_batched_tokens >= max_model_len "
            f"({self.config.max_model_len}); raising max_num_batched_tokens from "
            f"{self.config.max_num_batched_tokens} to {self.config.max_model_len}."
        )
        # 自动提升为用户兜底，用户显式配置的更大值不会被覆盖
        self.config.max_num_batched_tokens = self.config.max_model_len
```

tests/experimental/agent_loop/test_basic_agent_loop.py

适配 #7613 随机平局打破，修复三个 GPU CI 测试的 flaky 问题，是本次除引擎修复外的第二处关键改动。

```
def test_new_requests_route_to_least_loaded(self, ray_for_lb):
    lb = ray.remote(GlobalRequestLoadBalancer).remote(servers={"s0": None, "s1": None, "s2":
    None})
    # 第一次 acquire 可能命中任意副本：并列最少负载时平局为随机选择，
    # 因此先记录实际分配到的副本，再用同一个 request_id 堆积 inflight。
    s_heavy = ray.get(lb.acquire_server.remote(request_id="a"))[0]
```

```
ray.get(lb.acquire_server.remote(request_id="a"))[0]
ray.get(lb.acquire_server.remote(request_id="a"))[0]
# 第二个 request 必须落在另一副本上（其 inflight 为 1）
s_mid = ray.get(lb.acquire_server.remote(request_id="b"))[0]
assert s_mid != s_heavy
# 剩余副本 inflight 为 0，新请求应被路由到该空闲副本
idle = ({"s0", "s1", "s2"} - {s_heavy, s_mid}).pop()
s_new = ray.get(lb.acquire_server.remote(request_id="d"))[0]
assert s_new == idle
```

评论区精华

> wuxibin89：From #7629 #7624, I think we should provide max_model_len explicitly if enable_chunked_prefill=False：actor_rollout_ref.rollout.max_model_len="${MAX_MODEL_LEN}"。BTW，can we always enable_chunked_prefill=True for newer vllm version?
>
> ETOgaosion：I think we can make this enabled by default.

结论：两条路线都落地——引擎侧用自动提升兜底非 chunked 场景，示例与 E2E 默认走 chunked prefill；维护者对“显式传 max_model_len”的偏好通过脚本中保留 MAX_MODEL_LEN 推导而部分保留。

- 关闭 chunked prefill 时是否应显式传 max_model_len，以及新 vLLM 能否默认开启 (design): 采用引擎侧安全网自动提升 max_num_batched_tokens，同时示例与 E2E 回归默认 chunked prefill；显式 MAX_MODEL_LEN 推导在脚本中保留。
- 负载均衡测试在随机平局打破下的确定性 (testing): 更新三个测试，先记录实际分配再验证 least-loaded 语义，避免依赖 dict 顺序。

风险与影响

- 风险：
  1. 启动路径核心变更：vllm_async_server._validate_configs 是引擎启动必经路径，自动放大 max_num_batched_tokens 会提高单次调度迭代的上限，可能增加峰值显存 / KV cache 占用；仅影响非 chunked 场景，且用户显式配置更大值不会被覆盖，风险可控。
  2. 示例默认行为变化：GPU 示例从 enable_chunked_prefill=False 回归默认 True，若使用旧版 vLLM（仍受 VLM 占位符限制）可能复现 vllm#15185；该 issue 已关闭且在 vllm 0.24 下验证，风险较低但存在版本依赖。
  3. 测试确定性削弱：agent-loop 测试从“断言确定性分配”改为“观察后断言”，对随机平局打破了更健壮，但减弱了对异常路由行为的探测能力，平衡器行为回归时测试未必能捕获。
  4. 覆盖广但缺乏专项单测：安全网逻辑没有专门的单元测试，主要依赖 geo3k E2E 与 CI 验证；由于改动分散在 22 个文件，脚本清理若遗漏某处残留仍会静默影响对应工作流。- 影响：对用户：关闭 chunked prefill 的长上下文 vLLM 用户不再遭遇引擎启动崩溃，max_num_batched_tokens 行为与 vLLM 原生默认对齐；GPU 示例脚本恢复默认 chunked prefill，长 prompt 与 VLM 任务的吞吐和兼容性改善。对系统：geo3k VLM E2E 与 vLLM agent-loop CI 作业恢复确定性成功率，消除因默认参数不一致导致的“脚

本在启动阶段即失败”问题。对团队：确立了“引擎参数自愈 + 显式配置优先”的模式，可作为后续处理同类 vLLM 约束冲突的参考，也示范了在上游 issue 修复后及时清理 workaround 的节奏。 - 风险标记：启动路径核心变更，示例默认行为变化，测试确定性降低，缺少专项单测

关联脉络

- PR #7629 （未提供标题）：PR body 明确引用，作为移除 enable_chunked_prefill=False 覆盖项的上游相关 PR；本 PR 沿其方向继续清理示例与 E2E 脚本。
- PR #7624 （未提供标题，关联 issue）：PR body 与提交信息均引用，指出 vLLM 0.24 下 enable_chunked_prefill=False 与 max_num_batched_tokens 冲突导致的启动崩溃，是本 PR 的问题来源。
- PR #7613 （未提供标题）：改变了 load balancer 的平局打破行为，导致本 PR 必须同步更新 agent-loop 三个 GPU 测试。
- PR #7584 [ci] fix: drop stale enable_chunked_prefill=False from Ascend NPU scripts: 同属清理失效 enable_chunked_prefill=False 配置的系列工作，本 PR 处理 GPU 侧示例与 E2E。
- PR #7558 [ci] fix: drop stale enable_chunked_prefill=False from Ascend NPU scripts: 同类清理工作，证明该失效配置的清理正在跨 GPU/NPU 持续推进；本 PR 补上 GPU 示例与 CI 侧。