

PR #7654 完整报告

verl-project/verl

[ci] chore: fix ci failure with transformers==5.9.0

合并时间: 2026-09-01 21:03

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7654>

执行摘要

- 一句话: 修复 transformers 升级后 CI 失败, 更新模型和依赖配置。
- 推荐动作: 这是一个常规的 CI 修复 PR, 主要涉及配置调整和测试适配。值得快速浏览以了解 transformers 升级后的兼容性问题, 但无需深度精读核心逻辑。建议关注测试模型变更和依赖添加, 以确保后续类似升级时有参考。

功能与动机

本 PR 直接响应 PR#7650 (升级 transformers 到 5.9.0) 导致的 CI 失败。PR body 明确指出: 'Fix ci failure introduced in <https://github.com/verl-project/verl/pull/7650>'. 升级后, 多模态测试、SFT 引擎测试和 vLLM 相关测试出现不兼容, 需要调整配置和依赖。

实现拆解

1. 更新 CI 测试模型: 在多个 CI 工作流 (如 `e2e_ppo_trainer_megatron_sglang_2.yml`、`e2e_ppo_trainer_megatron_vllm_2.yml`) 中, 将模型 ID 从 `Qwen/Qwen2.5-VL-3B-Instruct` 改为 `Qwen/Qwen3.5-2B`, 以适配新版本 transformers。同时调整作业名称和依赖安装步骤。
2. 调整测试脚本: 在 `tests/special_e2e/sft/test_sft_engine_all.sh` 中, 注释掉 `veomni` 后端测试, 因为其尚未适配 transformers 5.9.0。在 `tests/experimental/agent_loop/test_multi_modal.py` 中, 添加视频支持条件检查 (基于环境变量), 并更新模型路径和断言逻辑以处理新版本的多模态处理。
3. 修复测试依赖: 在 `pyproject.toml` 的 `fsdp` 依赖中添加 `flash-linear-attention==0.5.2`, 并在 `uv.lock` 中更新锁文件。这是为了解决 transformers 5.9.0 可能引入的线性注意力相关依赖。
4. 修复 vLLM 测试: 在 `tests/workers/rollout/rollout_vllm/test_vllm_abort.py` 中, 为 `vLLMHttpServer` 实例添加缺失的属性初始化 (如 `_submission_paused`、`_admitting`、`_resume_event`), 以确保测试在新版本环境下正常运行。
5. CI 工作流清理: 在 `.github/workflows/vllm.yml` 中移除已注释的模型下载步骤, 并在 `.github/workflows/sgl.yml` 中添加测试路径触发条件。

关键文件:

- `tests/experimental/agent_loop/test_multi_modal.py` (模块 多模态测试; 类别 test; 类型 test-coverage): 核心测试文件, 包含实质性变更: 模型路径更新、视频支持条件化、断言

逻辑调整，以适配 transformers 5.9.0 的多模态处理行为。

- pyproject.toml (模块 依赖配置; 类别 config; 类型 configuration) : 依赖配置核心文件, 添加了 flash-linear-attention 依赖, 解决 transformers 5.9.0 升级后的兼容性问题。
- tests/special_e2e/sft/test_sft_engine_all.sh (模块 SFT 测试; 类别 test; 类型 test-coverage) : SFT 引擎测试脚本, 注释掉 veomni 后端测试, 反映版本适配的临时措施。
- .github/workflows/e2e_ppo_trainer_megatron_sglang_2.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure) : CI 工作流文件, 更新模型 ID 和作业名称, 确保 SGLang 和 Megatron 测试使用新模型。
- tests/workers/rollout/rollout_vllm/test_vllm_abort.py (模块 vLLM 测试; 类别 test; 类型 test-coverage) : vLLM 中止测试, 添加缺失的属性初始化, 修复测试在新环境下的运行时错误。

关键符号: 未识别

关键源码片段

tests/experimental/agent_loop/test_multi_modal.py

核心测试文件, 包含实质性变更: 模型路径更新、视频支持条件化、断言逻辑调整, 以适配 transformers 5.9.0 的多模态处理行为。

```
# 关键变更: 模型路径从 Qwen2.5-VL-3B-Instruct 更新为 Qwen3.5-2B, 以适配 transformers 5.9.0。
```

```
# 添加环境变量检查, 条件性包含视频提示, 因为 sglang 后端不支持视频。
```

```
VIDEO_PATH = os.path.expanduser("~/models/hf_data/test-videos/space_woaudio.mp4")
```

```
ROLLOUT_SUPPORTS_VIDEO = os.getenv("ROLLOUT_NAME", "vllm") != "sglang"
```

```
def init_config() -> DictConfig:
```

```
...
```

```
# 更新模型路径为 Qwen3.5-2B
```

```
model_path = os.path.expanduser("~/models/Qwen/Qwen3.5-2B")
```

```
config.actor_rollout_ref.model.path = model_path
```

```
...
```

```
def test_multimodal_tool_agent(init_config):
```

```
...
```

```
# 条件性构建视频提示列表
```

```
raw_prompts = [
```

```
    [{"role": "user", "content": "How are you?"}],
```

```
    [{"role": "user", "content": "Please generate a red image for me."}],
```

```
...
```

```
]
```

```
if ROLLOUT_SUPPORTS_VIDEO:
```

```
    raw_prompts.append([
```

```
        {
```

```
            "role": "user",
```

```
            "content": [
```

```
                {"type": "video", "video": VIDEO_PATH, ...},
```

```

        {"type": "text", "text": "Describe this video. ..."},
    ],
}
])
...
# 断言逻辑调整：根据多模态类型（视频或图像）检查轮次数
for i in range(len(num_turns)):
    multi_modal_type = parse_multi_modal_type(raw_prompts[i // n])
    if multi_modal_type == "video":
        assert num_turns[i] == 2 or num_turns[i] == 4, f"..."
    elif i // n == 0:
        assert num_turns[i] == 2, f"..."
    else:
        assert num_turns[i] == 4, f"..."

```

评论区精华

本 PR 无 review 讨论，因此没有争议点或决策结论。所有变更均为直接修复。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 依赖兼容性风险：添加 flash-linear-attention==0.5.2 可能引入与现有环境的冲突，尤其是当其他组件（如 Megatron 或 FSDP）有不同版本要求时。风险较低，因为这是 fsdp extra 的显式依赖。
 2. 测试覆盖风险：注释掉 veomni 测试可能暂时降低该后端的 CI 覆盖率，但注释表明这是因为版本不匹配，属于临时措施。
 3. 模型更新风险：将测试模型从 Qwen2.5-VL-3B-Instruct 改为 Qwen3.5-2B 可能影响多模态测试的准确性，但这是为适配新版本 transformers 的必要调整。
 4. CI 稳定性风险：更改 CI 工作流和作业名称可能影响依赖这些作业的其他流程，但变更范围限于命名和模型 ID，逻辑未变。- 影响：本 PR 主要影响 CI 流程和测试基础设施，不直接影响生产代码。用户层面无感知。团队需要关注 CI 通过情况，尤其是多模态和 SFT 测试。影响范围中等，因为涉及多个 CI 工作流和测试文件。- 风险标记：依赖版本兼容，测试覆盖临时降低

关联脉络

- PR #7650 [ci] chore: upgrade transformers==5.9.0: 本 PR 是 PR#7650 的直接后续修复，解决其引入的 CI 失败。