

PR #44349 完整报告

vllm-project/vllm

[Tests] Gate Step3VL under Transformers v5

合并时间: 2026-07-16 08:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44349>

执行摘要

- 一句话: 为 Step3VL 模型添加 Transformers v5.3 版本上限
- 推荐动作: 值得精读, 特别是对于参与 Transformers v5 升级的人员。此修复展示了如何处理上游 API 破坏性变更与 vLLM 内部 vendored 配置的关系, 并强调了正确设置 `check_version_reason` 键的重要性。

功能与动机

Transformers v5.4 移除了 `validate_rope()` 的 `ignore_keys` 参数, 导致 `Step3VLForConditionalGeneration` 的 HF runner 测试在 Transformers v5 上失败。此问题被跟踪在 issue #38379 中。PR 体中提到: “This is listed as an open item in the tracker issue #38379. This mirrors the identical fix already applied to `SarvamMLAForCausalLM` (same root cause, same fix pattern).”

实现拆解

变更仅涉及一个文件 `tests/models/registry.py`, 为 `Step3VLForConditionalGeneration` 的 `_HfExamplesInfo` 添加两个字段:

1. `max_transformers_version="5.3"`: 设定版本上限, 使测试在 Transformers >5.3 时跳过。
2. `transformers_version_reason`: 使用键 "hf" (而非最初使用的 "vllm"), 说明跳转原因。原因是 vLLM 已自行供应配置, 在 vLLM runner 下测试不受影响, 仅在 HF runner 下需要跳过。该设计确保 vLLM runner 测试在所有 Transformers 版本上继续运行, 而 HF runner 测试仅在 `<=5.3` 时运行。

关键文件:

- `tests/models/registry.py` (模块 测试注册表; 类别 test; 类型 test-coverage): 核心变更文件, 为 `Step3VLForConditionalGeneration` 添加版本上限和跳转原因。

关键符号: 未识别

关键源码片段

`tests/models/registry.py`

核心变更文件, 为 `Step3VLForConditionalGeneration` 添加版本上限和跳转原因。

```
# tests/models/registry.py
```

```
# 在模型注册字典中，为 Step3VLForConditionalGeneration 添加版本上限
"Step3VLForConditionalGeneration": _HfExamplesInfo(
    "stepfun-ai/step3",
    trust_remote_code=True,
    # Transformers v5.4 移除了 validate_rope() 的 ignore_keys 参数
    # vLLM 已自行供应此模型配置，因此不受影响
    max_transformers_version="5.3",
    transformers_version_reason={
        "hf": ( # 仅在 HF runner 下跳过，vLLM runner 不受影响
            "Transformers v5.4 removed the ignore_keys param from "
            "validate_rope(); vLLM has vendored the config and is unaffected"
        )
    },
),
```

评论区精华

审核者 hmellor 指出：“The skip reason is not vLLM. vLLM has vendored this config so it will work. This model will not work on Transformers anymore though.” 作者据此将原因键从 `"vllm"` 改为 `"hf"`，并详细解释了两种 runner 路径下版本检查行为的不同：vLLM runner 使用 `check_version_reason="vllm"` 不匹配 `"hf"` 键，因此版本上限无效，测试在所有版本运行；HF runner 使用默认 `check_version_reason="hf"` 匹配，从而正确跳过。

- 跳转原因键的选择：`"vllm"` vs `"hf"` (correctness): 将原因键改为 `"hf"`，使 vLLM runner 测试在所有版本运行，仅 HF runner 在 v5.4+ 跳过。

风险与影响

- 风险：风险极低。仅修改测试注册表中的元数据，不影响运行时或生成路径。版本上限可能导致未预期的跳过，但已通过手动验证确认为正确。
- 影响：影响范围仅限于测试执行：Step3VLForConditionalGeneration 的 HF runner 测试将在 Transformers v5.4+ 环境中跳过，避免因 API 变更导致的失败。vLLM runner 测试不受影响。对用户和使用者无影响。
- 风险标记：暂无

关联脉络

- PR #43305 Gate Tarsier2 under Transformers v5: 同一跟踪 issue #38379 的并行 PR，处理不同模型的类似问题。
- PR #38821 Gate Tarsier2 under Transformers v5: 同一跟踪 issue 的另一 PR，修复方式相同（加版本上限）。
- PR #44282 MiniCPMV LoRA test skip under Transformers v5: 同一跟踪 issue 的 PR，使用 `skipif` 跳过整个测试模块。