PR #42104 完整报告

vllm-project/vllm

[CI] set max transformers version for skywork model

合并时间：2026-05-14 07:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42104

执行摘要

- 一句话：限制 Skywork 模型 transformers 版本上限
- 推荐动作：作为临时修复，此 PR 快速解决了 CI 稳定问题。长期应关注 transformers 5.x 的兼容性，或推动 Skywork 官方修复其模型初始化。

功能与动机

修复因 transformers 5.x API 变更导致的 AttributeError: 'SkyworkChatModel' object has no attribute 'all_tied_weights_keys'，该错误使多模态测试（如 test_single_image_models 和 test_multi_image_models）在 AMD CI 上全部失败。问题源自 transformers 5.x 中 post_init() 会设置 all_tied_weights_keys，但 Skywork 模型未调用该方法。关联 Issue #42020 描述了该 CI 故障。

实现拆解

1. 定位问题：在 tests/models/registry.py 的 SkyworkR1VChatModel 注册条目中，最初只指定了模型名和 trust_remote_code=True，未限制 transformers 版本，导致在 transformers >=5.0 时因 API 不兼容而崩溃。
2. 添加版本上限：为 SkyworkR1VChatModel 注册项新增 max_transformers_version="4.57"，确保测试仅在 transformers <=4.57 时运行，避免新 API 冲突。
3. 补充原因说明：增加 transformers_version_reason 字典，键为 "hf"，详细解释为何需要上限：SkyworkChatModel.__init__ 未调用 self.post_init()，因此 all_tied_weights_keys 从未被设置；而 transformers v5 在 _move_missing_keys_from_meta_to_device 中要求此属性，导致错误。
4. 仅修改测试注册文件，不触及任何核心生产代码，确保变更风险极低。

关键文件：

- tests/models/registry.py（模块 测试注册；类别 test；类型 test-coverage）：唯一变更文件，为 SkyworkR1VChatModel 添加 transformers 版本上限及原因，防止在高版本 transformers 上因 API 不兼容导致 CI 测试失败。

关键符号：未识别

关键源码片段

tests/models/registry.py

唯一变更文件，为 SkyworkR1VChatModel 添加 transformers 版本上限及原因，防止在高版本 transformers 上因 API 不兼容导致 CI 测试失败。

```
# tests/models/registry.py
# SkyworkR1VChatModel 条目：限制 transformers 版本 <= 4.57
"SkyworkR1VChatModel": _HfExamplesInfo(
    "Skywork/Skywork-R1V-38B",
    trust_remote_code=True,
    max_transformers_version="4.57",
    # 注意：transformers_version_reason 中若包含 "hf" 键，
    # 会导致 check_transformers_version 无条件跳过 HF runner 测试。
    transformers_version_reason={
        "hf": (
            "SkyworkChatModel.__init__ 未调用 self.post_init()，"
            "因此 `all_tied_weights_keys` 从未被设置；"
            "Transformers v5 在 _move_missing_keys_from_meta_to_device 中要求此属性。"
        )
    },
),
```

评论区精华

gemini-code-assist[bot] 指出添加 transformers_version_reason（键为 "hf"）会导致测试被无条件跳过（即使版本满足上限），因为 check_transformers_version 函数中一旦该字典包含键，就会直接跳过测试。建议移除 transformers_version_reason 以确保 v4.x 测试仍可运行。但最终合并者 DarkLight1337 仍批准了该 PR，表明当时的意图可能是彻底跳过涉及 HF runner 的天空模型测试，原因可能是该模型依赖远程代码且维护困难。

- transformers_version_reason 导致测试无条件跳过 (testing): 尽管存在此问题，PR 仍被批准合并，表明当前意图是彻底跳过 Skywork 的 HF runner 测试。

风险与影响

- 风险：低风险。变更仅限于测试配置文件，无生产代码影响。唯一风险是若 transformers_version_reason 导致测试无条件跳过，可能遗漏后续兼容性回归。但鉴于该模型使用 trust_remote_code 且问题出在 HF 侧，跳过测试是可接受的临时措施。
- 影响：影响范围小。仅影响 Skywork 多模态测试在 transformers >=5.0 时的运行，避免 CI 持续失败。其他模型不受影响。
- 风险标记：测试覆盖未完全验证

关联脉络

- PR #42020 [CI Failure]: mi300_1: Multi-Modal Models (Extended Generation 2): 该 Issue 记录了 Skywork 多模态测试在 AMD CI 上的失败，是此 PR 的直接修复目标。
- PR #42536 Remove verifier model type check in speculative config: 虽无直接关联，但同样涉及模型注册配置调整，属于同一维护类别。