

PR #44239 完整报告

vllm-project/vllm

[Bugfix][CI/Build] Fix Plamo2 HF runner crash on transformers v5 (`_tied_weights_keys` list→dict)

合并时间: 2026-07-23 20:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44239>

执行摘要

- 一句话: 修复 Plamo2 在 transformers v5 下的 HF runner 崩溃
- 推荐动作: 可快速合入。该 PR 是 Transformers v5 升级跟踪的一部分, 修复方式合理, 逻辑清晰, 可作为未来类似兼容问题的修复模板。

功能与动机

关联 Issue #38379 (Transformers v5 升级跟踪) 中记录了 Plamo2 在 transformers v5 下的测试失败。PR body 指出, `HfRunner.from_pretrained()` 在 transformers v5 中因 `_tied_weights_keys` 类型不匹配而抛出 `AttributeError: 'list' object has no attribute 'keys'`。生产代码不受影响 (vLLM 通过 `config.tie_word_embeddings` 标志处理权重绑定), 但测试基础设施不可用。

实现拆解

1. 在 `tests/conftest.py` 新增 `_fix_v4_tied_weights_keys` 函数: 该函数接收一个模型类, 检查其 `_tied_weights_keys` 是否为非空 list, 若是则转换为 `dict[str, str]` 格式 (将包含 "lm_head" 且以 ".weight" 结尾的键映射到 "model.embed_tokens.weight"), 并通过 `setattr` 就地修改类属性, 该修改通过 `sys.modules` 缓存保留到 `from_pretrained()` 中。
2. 在 `HfRunner.__init__()` 中集成修复: 在调用 `auto_cls.from_pretrained()` 之前, 通过 `try_get_class_from_dynamic_module` 获取模型类, 并应用 `_fix_v4_tied_weights_keys`。只处理当前 `auto_cls` 对应的 `auto_map` 条目, 避免影响其他模型。
3. 在 `tests/models/registry.py` 中移除 Plamo2 的版本限制: 删除 `max_transformers_version="4.57"` 及对应的原因说明, 使 Plamo2 能在 transformers v5 下参与测试。

关键文件:

- `tests/conftest.py` (模块 测试工具; 类别 test; 类型 test-coverage; 符号 `_fix_v4_tied_weights_keys`): 新增 `_fix_v4_tied_weights_keys` 函数并在 `HfRunner.__init__()` 中调用, 是修复的核心逻辑。
- `tests/models/registry.py` (模块 测试注册; 类别 test; 类型 test-coverage): 移除 Plamo2 的 `max_transformers_version` 限制, 使测试在 transformers v5 下运行。

关键符号: `_fix_v4_tied_weights_keys`


```
if model_cls is not None:
    _fix_v4_tied_weights_keys(model_cls)

model = cast(nn.Module, auto_cls.from_pretrained(...))
```

评论区精华

该 PR 没有 review 评论，但有 7 个提交的演进历史体现了关键决策：

- 最初实现使用 `auto_map` 所有条目迭代 + 手动 `warn_on_fail`，后改为使用 vLLM 已有的 `try_get_class_from_dynamic_module` 函数。
- 曾尝试直接删除 `_tied_weights_keys`，验证后发现会导致权重重新初始化为随机值，模型输出错误。
- 曾添加单元测试文件 `test_fix_v4_tied_weights_keys.py`，但因 vLLM 无测试 `conftest` 工具的先例而被移除。
- 暂无高价值评论线程

风险与影响

- 风险：变更仅限测试基础设施，不影响 vLLM 生产代码。风险极低：
 1. `_fix_v4_tied_weights_keys` 仅当 `_tied_weights_keys` 为 list 且包含 "lm_head" 时才生效，其他模型不受影响。
 2. 使用 `try_get_class_from_dynamic_module` 失败时返回 None，不会引发异常。
 3. 删除版本限制后，Plamo2 测试在 transformers v5 下可能因其他依赖（如 `causal_conv1d` / `mamba_ssm`）而失败，但属于已有问题，与本次变更无关。- 影响：影响范围小：仅改变测试基础架构，使 Plamo2 模型能在 transformers v5 环境中运行测试。对其他模型或生产部署无影响。- 风险标记：暂无

关联脉络

- PR #38379 Upgrade to Transformers v5: 本 PR 是 #38379 的子任务，专门修复 Plamo2 在 transformers v5 下的兼容问题。
- PR #44282 PR #44282: 同样是 #38379 的子任务，处理 MiniCPMV 等模型的 transformers v5 兼容性。