

# PR #40389 完整报告

vllm-project/vllm

Deprecate support for Transformers v4

合并时间: 2026-04-27 23:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/40389>

## 执行摘要

- 一句话: 废弃 Transformers v4 兼容代码路径
- 推荐动作: 值得快速阅读, 了解项目的版本依赖弃用策略和 review 讨论中关于时间线妥协的决策过程。

## 功能与动机

根据 PR body 说明: "As per deprecation policy." —— 即按照 vLLM 项目的弃用策略, 需要正式宣布停止支持 Transformers v4, 并引导用户升级到 v5。

## 实现拆解

1. 版本检查与日志输出: 在 `vllm/transformers_utils/config.py` 的模块加载阶段 (`logger = init_logger(__name__)` 之后), 新增一段条件判断: 使用 `from packaging.version import Version` 和 `from importlib.metadata import version` 获取已安装的 transformers 版本, 若低于 5.0.0, 则通过 `logger.warning()` 输出弃用信息。
2. 弃用时间线: 根据 reviewer DarkLight1337 的建议, 将警告信息分为两个阶段: v0.22.0 起停止维护 v4 代码路径, v0.24.0 彻底移除。
3. 移除多余 import: 初始版本使用了 `import warnings` 和 `warnings.warn()`, 但 `gemini-code-assist[bot]` 指出 `DeprecationWarning` 默认被隐藏, 建议改用 `logger.warning()`。最终实现采纳了该建议, 因此 `import warnings` 被移除。

关键文件:

- `vllm/transformers_utils/config.py` (模块 配置模块; 类别 `source`; 类型 `core-logic`): 这是本 PR 唯一变更的文件, 新增 Transformers v4 弃用警告。

关键符号: 未识别

## 关键源码片段

`vllm/transformers_utils/config.py`

这是本 PR 唯一变更的文件, 新增 Transformers v4 弃用警告。

```
# vllm/transformers_utils/config.py
# ... 前置 import 省略 ...

logger = init_logger(__name__)
```

```
# 弃用 Transformers v4: 检查版本并发出警告
if Version(version("transformers")) < Version("5.0.0"):
    logger.warning(
        "Support for Transformers v4 is deprecated. The Transformers v4 codepath will "
        "become unmaintained in vLLM v0.22.0 and will be removed in vLLM v0.24.0. "
        "Please upgrade to Transformers v5: pip install --upgrade transformers"
    )

# 后续代码不变, 包括 LazyConfigDict 定义、_CONFIG_REGISTRY 等
```

## 评论区精华

1. 使用 `logger` 替代 `warnings`: `gemini-code-assist[bot]` 指出 `DeprecationWarning` 默认隐藏, 用户不易看到, 建议改用 `logger.warning()` 以保可见性, 并与现有废弃模式一致。该建议被采纳, 最终代码中去掉了 `import warnings`。
  2. 弃用时间窗口: `DarkLight1337` 建议给予更长的过渡时间, 将完全移除从 `v0.22.0` 推迟到 `v0.24.0`, 并在 `v0.22.0` 起停止主动维护。`hmellor` 最初有顾虑, 但最终同意该方案。
- 使用 `logger.warning()` 替代 `warnings.warn()` (`correctness`): 采纳建议, 最终版本使用 `logger.warning()`, 去掉了 `import warnings`。
  - 弃用时间窗口的协商 (`design`): 警告内容调整为: `v0.22.0` 停止维护, `v0.24.0` 彻底移除。

## 风险与影响

- 风险: 本 PR 仅引入版本检查与日志输出, 未修改任何现有功能逻辑, 因此回归风险极低。但需注意: 如果未来在 `v0.24.0` 实际移除 `v4` 代码路径, 可能影响仍在 `使用 Transformers v4` 的用户, 届时需要确保所有相关路径已迁移或提供替代方案。
- 影响: 对用户: 所有仍在 `使用 Transformers v4` 的用户将在启动时看到显式的黄色警告, 提示升级。对系统: 无功能影响, 仅增加一次版本检查和一行日志输出, 性能开销可忽略。对团队: 确立了 `v0.22.0` 停止维护、`v0.24.0` 移除的时间线, 便于后续规划。
- 风险标记: 暂无

## 关联脉络

- 暂无明显关联 PR