

PR #52431 完整报告

vllm-project/vllm

[Bugfix] Fix modelscope usage

合并时间: 2026-08-15 17:40

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52431>

执行摘要

- 一句话: 修复 ModelScope KeyError 并支持 VLLM_USE_MODELSCOPE=1
- 推荐动作: 值得快速浏览, 特别是使用 ModelScope 的团队; 设计上强调与 Hugging Face Hub 行为对齐和布尔环境变量的一致性。代码简单, 无深层架构影响。可后续补充一个针对 modelscope_list_repo_files 的 mock 单元测试, 防止第三方 API 变更再次破坏。

功能与动机

PR body 明确指出: 调用 modelscope_list_repo_files 时, 新版 ModelScope 不再返回 Type 字段, 导致 KeyError (作者在通过 ModelScope 加载 meta-models/Muse-Glimmer-30B 时遇到); 同时 VLLM_USE_MODELSCOPE=1 未被识别, 与其他布尔环境变量的用法不一致, 因此需要修复。

实现拆解

1. 修改 vllm/transformers_utils/utils.py 中 modelscope_list_repo_files: 移除 file["Type"] == "blob" 过滤条件, 直接返回所有 file["Path"]。原因是新版 ModelScope 的 get_model_files 返回对象不再包含 Type 键, 原有过滤会触发 KeyError; recursive=True 时返回的项目均为文件, 过滤本身已无必要。
2. 修改 vllm/envs.py 中 VLLM_USE_MODELSCOPE 解析逻辑: 从 .lower() == "true" 改为 .strip().lower() in ("1", "true"), 同时删除注释中 not numbers 的说明, 接受 1 作为开启值, 与 VLLM_USE_FASTTOKENS 等布尔型环境变量风格对齐。
3. 无测试或文档配套修改; PR body 的 Test Plan 为空, 已触发 Buildkite CI (#84016)。

关键文件:

- vllm/transformers_utils/utils.py (模块 模型工具; 类别 source; 类型 core-logic; 符号 modelscope_list_repo_files): 核心修复: modelscope_list_repo_files 移除对 file["Type"] 的依赖, 消除新版 ModelScope API 导致的 KeyError, 同时简化返回逻辑。
- vllm/envs.py (模块 环境配置; 类别 source; 类型 core-logic): 让 VLLM_USE_MODELSCOPE=1 生效, 与其他布尔环境变量解析风格对齐, 降低用户配置成本。

关键符号: modelscope_list_repo_files

关键源码片段

vllm/transformers_utils/utils.py

核心修复：modelscope_list_repo_files 移除对 file["Type"] 的依赖，消除新版 ModelScope API 导致的 KeyError，同时简化返回逻辑。

```
def modelscope_list_repo_files(
    repo_id: str,
    revision: str | None = None,
    token: str | bool | None = None,
) -> list[str]:
    """List files in a modelscope repo."""
    from modelscope.hub.api import HubApi

    api = HubApi()
    api.login(token)

    # 与 huggingface_hub.list_repo_files 对齐：
    # 新版本 ModelScope 返回的文件对象不再包含 "Type" 键，
    # 若继续按 file["Type"] == "blob" 过滤会触发 KeyError，
    # 因此这里直接收集所有文件的 "Path" 字段。
    return [
        file["Path"]
        for file in api.get_model_files(
            model_id=repo_id, revision=revision, recursive=True
        )
    ]
```

评论区精华

三位维护者（noooooop、Isotr0py、jeejeelee）批准了该 PR，无实质性 review 讨论。

claude[bot] 自动评论指出 fork PR 禁用自动 review，可手动触发。Issue 评论仅为 CI 触发与确认，无技术交锋。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 移除 Type == "blob" 过滤后，若未来 ModelScope API 在 recursive=True 时返回非文件对象（如目录标记），可能将非文件路径混入结果，导致后续下载或加载失败。当前假设 API 只返回文件。
 - VLLM_USE_MODELSCOPE 解析改为接受 "1"，与常见布尔环境变量一致；但 "2" 等不被接受，行为符合预期。
 - 缺少针对 modelscope_list_repo_files 的单元测试，未来第三方 API 改动可能再次引入回归。
- 影响：

- 用户影响：修复依赖 ModelScope 下载模型的用户（尤其中国大陆用户）在加载模型时的 `KeyError`，并允许 `VLLM_USE_MODELSCOPE=1` 这一常见写法。
- 系统影响：仅涉及模型仓库枚举与启动配置解析，不改变核心推理路径；但 `env` 行为的变化对所有使用 `VLLM_USE_MODELSCOPE` 的部署配置生效。
- 团队影响：改动小、无 API 变更，风险低。
- 风险标记：第三方 API 兼容性依赖，缺少测试覆盖

关联脉络

- PR #52289 [Doc] Update model support information: PR body 提到通过 ModelScope 加载 `meta-models/Muse-Glimmer-30B`，与 #52289 登记的 Muse Glimmer 模型支持直接相关。