PR #39292 完整报告

vllm-project/vllm

[CI Failure] pin nomic-embed-text-v1 revision

合并时间：2026-04-08 19:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39292

执行摘要

本 PR 通过 pin nomic-embed-text-v1 模型的特定 revision，修复了因模型更新支持 transformers v5 导致的 CI 测试失败。变更仅限于测试代码，是临时解决方案，旨在恢复 CI 稳定性，对用户无直接影响。

功能与动机

动机源于 nomic-ai/nomic-embed-text-v1 模型更新支持 transformers v5 后引发 CI 失败。PR body 明确引用 buildkite 日志，指出：'nomic-ai/nomic-embed-text-v1 was just updated to support transformers v5, but it caused CI failure.' 目的是快速修复 CI 中断，确保测试通过。

实现拆解

实现围绕测试代码调整展开：

- 核心数据结构变更：在 tests/models/utils.py 的 ModelInfo 类中添加 revision 字段，使其可选，支持传递模型版本信息。
- 测试调用更新：在多个测试文件中更新 vllm_runner 和 hf_runner 调用，传递 revision 参数。例如，在 tests/models/language/pooling/test_nomic_max_model_len.py 中： with vllm_runner(model_info.name, revision=model_info.revision, # 新增参数 runner="pooling", max_model_len=None)
- 特定模型修复：为 nomic-embed-text-v1 设置 revision="720244025c1a7e15661a174c63cce63c8218e52b"，并添加 Fixme 注释提醒未来移除。

关键文件如下： | 文件路径 | 修改内容 | 重要性 | |----------|----------|--------| | tests/models/utils.py | 添加 revision 字段到 ModelInfo | 高，数据结构基础 | | tests/models/language/pooling/test_nomic_max_model_len.py | 更新多个测试调用传递 revision | 中，直接影响测试执行 | | tests/models/language/pooling_mteb_test/test_nomic.py | 为 nomic 模型设置特定 revision | 低，特定模型修复 |

评论区精华

review 中无深度讨论，仅 gemini-code-assist[bot] 确认更改一致实现所需功能，DarkLight1337 批准。这表明团队对修复方案无争议，快速推进以解决 CI 问题。

> gemini-code-assist[bot]: "I have no feedback to provide as the changes consistently implement the required revision tracking across the test suite."

风险与影响

- 技术风险：pin revision 是临时修复，可能导致测试落后于模型最新版本，长期引发兼容性问题。Fixme 注释提示未来更新，但缺乏时间表，存在维护债务。修改涉及多个测试文件，需确保所有调用点正确传递 revision，否则遗漏可能导致测试失败。
- 影响分析：对用户无影响，变更限于测试代码。对系统恢复 CI 稳定性，防止构建失败。对团队减少 CI 中断，但增加后续更新负担，影响程度小。

关联脉络

与历史 PR 关联显示 CI 和测试修复的常见模式：

- PR 37025：添加推理解析器测试到 CI，共享关注测试稳定性和 CI 集成。
- PR 39284：修复文档构建崩溃，类似解决外部依赖更新导致的构建问题。

这些 PR 共同反映 vLLM 项目在快速演进中，需频繁调整测试以适应外部模型更新，强调了 CI 维护的重要性。