PR #39176 完整报告

vllm-project/vllm

fix(test): recompute Jina ColBERT rotary inv_freq cleared by transformers v5 weight loader

合并时间：2026-04-07 22:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39176

执行摘要

- 一句话：修复 Transformers v5 权重加载后 Jina ColBERT 模型旋转嵌入 inv_freq 缓冲区被清空导致的 NaN 输出问题。
- 推荐动作：该 PR 值得快速浏览，了解 Transformers v5 权重加载机制对非持久化缓冲区的影响。关注点：1）非持久化缓冲区在权重加载中的处理变化；2）测试中模型状态恢复的模式。

功能与动机

修复 Transformers v5 中 Jina ColBERT 模型测试失败问题。根据 PR body 和关联 Issue #38737，Transformers v5 的新权重物化系统会清除非持久化缓冲区，而 Jina ColBERT 模型的 RotaryEmbedding 模块的 inv_freq 缓冲区被注册为 persistent=False，导致加载后被清空，使模型产生 NaN 输出，测试 test_colbert_hf_comparison[jina] 失败。

实现拆解

在测试文件 tests/models/language/pooling/test_colbert.py 的 _load_hf_model 函数中，模型加载并移动到设备后，添加代码遍历所有模块，检查是否具有 _compute_inv_freq 方法和 inv_freq 属性，然后重新计算 inv_freq 缓冲区。这确保了模型在权重加载后能产生有效输出。

关键文件：

- tests/models/language/pooling/test_colbert.py（模块 测试 / 模型）：唯一修改的文件，包含修复逻辑，直接影响测试通过性。

关键符号：_load_hf_model

评论区精华

review 中 gemini-code-assist[bot] 指出原始实现使用 mod.inv_freq.device 进行重新计算可能脆弱，因为缓冲区被清空后其元数据可能不一致或为 None。建议直接使用函数作用域中已有的 device 变量，更安全可靠。作者 ieBoytsov 接受了此建议并修改了代码。

- 重新计算 inv_freq 时设备变量的使用 (correctness): 作者接受建议，修改代码使用 device 变量。

风险与影响

- 风险：风险较低，主要影响测试环境。变更仅针对测试文件中的模型加载函数，不涉及生产代码。潜在风险包括：1）对非旋转嵌入模块的误操作（但通过 hasattr 检查避免）；2）如果其他模型有类似非持久化缓冲区问题，此修复可能不全面；3）依赖 _compute_inv_freq 方法的存在性。
- 影响：影响范围有限，主要修复特定测试用例。对用户无直接影响，仅确保测试通过。对系统无性能或功能影响。对团队而言，解决了 Transformers v5 升级导致的测试失败问题，有助于维护测试稳定性。
- 风险标记：测试环境变更，依赖特定方法存在性

关联脉络

- PR #37292 Fix Mistral yarn warning in Transformers v5: 同样处理 Transformers v5 兼容性问题，涉及版本检查和配置设置。
- PR #38763 only patch runtime_env for torch >= 2.10: 类似版本兼容性修复，关注依赖库版本对行为的影响。