

PR #49269 完整报告

vllm-project/vllm

Update BGE-M3 token expectations for leading spaces

合并时间: 2026-07-21 11:49

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49269>

执行摘要

- 一句话: 修复 BGE-M3 测试中 token 前导空格期望值
- 推荐动作: 建议快速合并以解封 CI, 这是对上游变更的被动适配, 无设计深度。测试维护者应关注 SentencePiece 行为变化是否稳定, 并考虑增加更全面的测试输入覆盖。

功能与动机

PR #48674 修改了 BGE-M3 稀疏嵌入插件的行为, 使 SentencePiece 保留前导空格 (如 `__What` 解码为 `" What"` 而非 `"What"`), 导致现有测试中的 token 期望值与实际输出不匹配, CI 失败。此 PR 旨在更新测试期望值以对齐改动, 快速解封 CI。

实现拆解

1. 定位问题: PR #48674 恢复了 SentencePiece 的前导空格, 使得 token 字段包含前导空格 (如 `" What"`), 而测试硬编码了无空格的 token 字符串 (如 `"What"`), 导致断言失败。
2. 修改测试期望值: 在 `tests/plugins_tests/test_bge_m3_sparse_io_processor_plugins.py` 的 `_check_sparse_embedding` 函数中, 将 `expected_weights` 列表内所有 token 字符串前面加上空格, 共 6 处修改 (+6/-6)。例如 `"token": "the"` 改为 `"token": " the"`, `"token": "What"` 改为 `"token": " What"` 等。
3. 保持其他逻辑不变: 测试的权重值、`token_id`、断言逻辑和 `check_tokens` 路径均未修改, 仅调整 token 字符串以匹配插件实际输出。
4. CI 验证: 运行 `ruff-format` 和 `ruff-check` 确保代码风格合规, 并通过 `git diff --check` 检查无残留差异。

关键文件:

- `tests/plugins_tests/test_bge_m3_sparse_io_processor_plugins.py` (模块测试; 类别 `test`; 类型 `test-coverage`; 符号 `_check_sparse_embedding`): 唯一变更文件, 调整了 BGE-M3 稀疏嵌入插件测试中的 token 期望值以适配前导空格行为, 直接解除 CI 阻塞。

关键符号: `_check_sparse_embedding`

关键源码片段

`tests/plugins_tests/test_bge_m3_sparse_io_processor_plugins.py`

唯一变更文件，调整了 BGE-M3 稀疏嵌入插件测试中的 token 期望值以适配前导空格行为，直接解除 CI 阻塞。

```
# tests/plugins_tests/test_bge_m3_sparse_io_processor_plugins.py

def _check_sparse_embedding(data, check_tokens=False):
    expected_weights = [
        {"token_id": 32, "weight": 0.0552978515625, "token": "?"},
        # 以下 token 字符串均添加了前导空格，对齐 SentencePiece 解码行为 (#48674)
        {"token_id": 70, "weight": 0.09808349609375, "token": " the"},
        {"token_id": 83, "weight": 0.08154296875, "token": " is"},
        {"token_id": 111, "weight": 0.11810302734375, "token": " of"},
        {"token_id": 4865, "weight": 0.1171875, "token": " What"},
        {"token_id": 9942, "weight": 0.292236328125, "token": " France"},
        {"token_id": 10323, "weight": 0.2802734375, "token": " capital"},
    ]
    expected_embed = {x["token_id"]: x for x in expected_weights}
    # ... 其余断言逻辑不变
```

评论区精华

无 review 评论线程。PR 由 noooop 批准合并，主要讨论在 Issue 评论中：noooop 询问是否应先合并以解封 CI，staugust 同意合并，并表示若需保持之前输出可后续调整插件。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅修改测试硬编码的期望值，不涉及生产代码逻辑。权重和 token_id 均未改动，断言函数相同，因此回归风险极小。但需注意：若未来 SentencePiece 行为再次变化，测试需同步更新；当前测试仅覆盖了特定输入 ("What is the capital of France?")，可能无法覆盖所有变化场景。
- 影响：影响范围小。仅影响 BGE-M3 稀疏嵌入插件的测试文件，使该测试与最新插件行为一致，从而解除 CI 阻塞。对生产系统、用户接口和其他模块无影响。
- 风险标记：依赖上游行为变更，仅覆盖单输入场景

关联脉络

- PR #48674 Restore leading spaces in BGE-M3 sparse plugin: 本 PR 的所有测试期望值更新正是为了适配 PR #48674 恢复的前导空格行为，两者直接关联。