

PR #46891 完整报告

vllm-project/vllm

[ROCm][CI] Add TRITON_ATTN score absolute tolerance floor

合并时间: 2026-06-27 14:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46891>

执行摘要

- 一句话: 为 TRITON_ATTN 后端添加绝对容忍度下限
- 推荐动作: 该 PR 属于典型的测试稳定性修复, 技术含量低但实用。值得关注其处理小概率值漂移的策略——为相对容忍度设置绝对下限——在类似浮点精度测试中可复用。

功能与动机

在 ROCm 环境下, TRITON_ATTN 后端的 text_vs_text 分数 (~0.10) 存在约 0.008 的固定绝对漂移, 虽然其他较大分数 (0.53, 0.74) 仍在相对容忍度内, 但小概率值的相对误差高达 7.9%, 超过了原有的 0.045 相对容忍度。ROCM_AITER_FA 和 FLEX_ATTENTION 后端已有类似的绝对容忍度下限 (分别来自 PR#41341), 而 TRITON_ATTN 是唯一缺少此容差的高漂移后端。

实现拆解

仅修改了一处配置: 在 `tests/entrypoints/pooling/scoring/test_cross_encoder_online_vision.py` 中的 `BACKEND_ABS_TOL` 字典里新增了 `"TRITON_ATTN": 0.009` 条目。同时调整了相关注释, 以说明绝对漂移在不同后端间的普遍性和影响。

关键文件:

- `tests/entrypoints/pooling/scoring/test_cross_encoder_online_vision.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `BACKEND_ABS_TOL`, `get_abs_tol`, `assert_score`): 单文件变更, 为 TRITON_ATTN 后端添加绝对容忍度下限 0.009, 并更新注释说明绝对漂移的普遍性。

关键符号: `get_abs_tol`, `assert_score`

关键源码片段

`tests/entrypoints/pooling/scoring/test_cross_encoder_online_vision.py`

单文件变更, 为 TRITON_ATTN 后端添加绝对容忍度下限 0.009, 并更新注释说明绝对漂移的普遍性。

```
# tests/entrypoints/pooling/scoring/test_cross_encoder_online_vision.py
```

```
# 部分 ROCm 注意力后端在低 text-vs-text 概率上显示小的绝对漂移,
```

```
# 即使较大的分数仍在相对容忍度内。绝对漂移在各个分数大小上一致
# (~0.005-0.010) , 因此它只对较小的 ~0.10 text-vs-text 值超过
# 相对容忍度。保持相对容忍度严格, 只为受影响的后端添加小的绝对下限。
# TRITON_ATTN: gfx942/ROCm 7.2 上 text-vs-text 漂移 ~0.008 绝对值
# (~7.9% 相对) 。
BACKEND_ABS_TOL: dict[str, float] = {
    "default": 0.0,
    "ROCM_AITER_FA": 0.005,
    "TRITON_ATTN": 0.009, # 新增: 解决 ~0.008 绝对漂移
    "FLEX_ATTENTION": 0.006,
}
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低: 仅调整了测试容忍度配置, 不影响任何生产逻辑。0.009 的绝对容忍度与 ROCM_AITER_FA (0.005) 和 FLEX_ATTENTION (0.006) 处于同一量级, 且仅作用于小概率值, 不会掩盖其他分数的大误差。
- 影响: 仅影响 ROCm 平台下 TRITON_ATTN 后端的 cross-encoder 在线视觉测试, 消除一次假阳性失败。对其他平台或其他后端无影响。
- 风险标记: 暂无

关联脉络

- PR #41341 Add absolute tolerance for ROCM_AITER_FA and FLEX_ATTENTION: 该 PR 为 ROCM_AITER_FA 和 FLEX_ATTENTION 添加了绝对容忍度下限, 本 PR 为 TRITON_ATTN 做了相同的处理。