

PR #50917 完整报告

vllm-project/vllm

[ROCm][Test] Use BF16 for Jina v5 nano MTEB test

合并时间: 2026-08-04 08:33

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50917>

执行摘要

- 一句话: Jina v5 nano MTEB 测试在 ROCm 上改用 BF16
- 推荐动作: 该 PR 值得快速合入, 虽然改动很小, 但它解决了 AMD CI 的实际失败, 并展示了按平台差异化设置测试精度的模式。关注点: 平台条件 dtype 的写法可复用, 后续类似的精度敏感性测试可参考。无需精读。

功能与动机

PR body 明确指出: AMD Buildkite 构建 11591 失败, 新加入的 Jina nano 用例暴露了长上下文 FP16 在 MI250 上的数值漂移。作者希望 'Run jinaai/jina-embeddings-v5-text-nano in BF16 only on ROCm while retaining the existing automatic FP16 selection on NVIDIA', 并保留 Hugging Face FP32 基线和长上下文嵌入比较作为正确性校验。

实现拆解

1. 在测试文件 tests/models/language/pooling_mteb_test/test_jina.py 的导入区新增 from vllm.platforms import current_platform, 用于在运行时获取当前加速器平台。
2. 为 jinaai/jina-embeddings-v5-text-nano 的 EmbedModelInfo 配置新增 dtype 字段, 值由三元表达式 "bfloat16" if current_platform.is_rocm() else "auto" 决定: ROCm 上强制使用 bfloat16, 避免 FP16 长上下文数值漂移; 其他平台 (如 NVIDIA) 保持 auto 自动选择 (通常为 FP16), 不改变原有覆盖。
3. 该条件仅在模型信息构造时求值一次, 不影响测试其余逻辑。测试仍通过 mteb_test_embed_models 执行 Hugging Face 与 vLLM 的对比, FP32 基线和长上下文检查保持激活, 没有用缓存 MTEB 分数绕过。
4. 无生产代码、配置或部署改动, 纯测试配置调整。

关键文件:

- tests/models/language/pooling_mteb_test/test_jina.py (模块 Jina 测试; 类别 test; 类型 test-coverage): 唯一变更文件, 为 Jina v5 nano 测试增加 ROCm 平台下的 BF16 dtype 配置, 修复 MI250 长上下文 FP16 漂移导致的 CI 失败。

关键符号: 未识别

关键源码片段

tests/models/language/pooling_mteb_test/test_jina.py

唯一变更文件，为 Jina v5 nano 测试增加 ROCm 平台下的 BF16 dtype 配置，修复 MI250 长上下文 FP16 漂移导致的 CI 失败。

```
# tests/models/language/pooling_mteb_test/test_jina.py
from vllm.platforms import current_platform

# 为 Jina v5 nano 模型按平台选择 dtype:
# ROCm 上使用 bfloat16 规避 MI250 长上下文 FP16 数值漂移;
# NVIDIA 等平台保持 "auto" 自动选择 (通常为 FP16), 保留原覆盖。
EmbedModelInfo(
    "jinaai/jina-embeddings-v5-text-nano",
    architecture="JinaEmbeddingsV5Model",
    dtype="bfloat16" if current_platform.is_rocm() else "auto",
    seq_pooling_type="LAST",
    attn_type="encoder_only",
    is_prefix_caching_supported=False,
    is_chunked_prefill_supported=False,
)
```

评论区精华

该 PR 没有实质性 review 讨论。claude[bot] 仅输出仓库配置的自动提示 (要求人工 review)，nooooop 直接批准 (state: APPROVED, 无评论)。因此没有争议点或未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，改动仅涉及测试文件。潜在风险点：1) current_platform.is_rocm() 的平台判定若在特定环境出现误判，可能导致 dtype 选择不符合预期，但该 API 在 vLLM 中广泛使用，可靠性高；2) ROCm 上强制 bfloat16 可能掩盖其他精度问题，但本用例旨在规避已发现的 FP16 漂移，且 HF 基线对比仍在；3) 改动只影响 jina-embeddings-v5-text-nano 一个模型用例，其他 Jina 模型和 reranker 测试不受影响。
- 影响：影响范围限定为 ROCm CI 中的 Jina nano 嵌入测试。改造后 AMD Buildkite 构建不再因 FP16 数值漂移失败，同时 NVIDIA 平台保持原有 FP16 覆盖不变。对用户和产品无直接影响，属于测试基础设施的稳定性改进。
- 风险标记：测试仅平台适配，依赖平台检测

关联脉络

- 暂无明显关联 PR