# PR #38508 完整报告

vllm-project/vllm

[ROCm][CI] Fix Whisper translation test attention backend selection

合并时间：2026-03-31 13:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38508

## 执行摘要

本 PR 修复了 ROCm 平台上 Whisper 翻译测试因注意力后端选择错误而失败的问题，通过动态配置函数确保使用兼容后端，提升 CI 稳定性。

## 功能与动机

该 PR 旨在解决 Whisper 模型在 ROCm 平台测试中使用 ROCM_AITER_FA 后端失败的问题。根据 PR body 描述，"ROCM_AITER_FA does not support ENCODER_DECODER cross-attention"，因此需要为 Whisper 选择支持交叉注意力的后端，如 ROCM_AITER_UNIFIED_ATTN 或 TRITON_ATTN。

## 实现拆解

实现集中在 tests/entrypoints/openai/speech_to_text/test_translation_validation.py 文件中。核心新增函数 _get_rocm_attention_config 动态选择注意力后端：

```
def _get_rocm_attention_config(model_name):
    if not current_platform.is_rocm():
        return None
    if "whisper" in model_name.lower():
        try:
            from vllm.platforms.rocm import _ON_MI3XX
            if _ON_MI3XX:
                return {"backend": "ROCM_AITER_UNIFIED_ATTN"}
        except ImportError:
            logger.warning(...)
            return {"backend": "TRITON_ATTN"}
    return {"backend": "ROCM_AITER_FA"}
```

更新了 server fixture 和测试函数 test_non_asr_model、test_basic_audio_with_lora，移除对 rocm_aiter_fa_attention fixture 的依赖，改为调用 _get_rocm_attention_config 获取配置。

## 评论区精华

review 讨论中突出两个要点：

- 模型名检查范围：gemini-code-assist[bot] 建议扩展检查到 granite-speech 模型，但作者 AndreasKaratzas 回应 "Granite-speech is not based on the Whisper architecture and has no CrossAttention"，因此保持逻辑不变。
- 异常处理具体化：gemini-code-assist[bot] 建议用 ImportError 替代 Exception，作者采纳并添加日志，回应 "Good suggestion, I also added a log for that instead of just passing :)"。

## 风险与影响

风险：

- 模型名检查可能不完全覆盖其他类似语音模型，导致后端选择错误。
- 平台检测依赖 current_platform.is_rocm() 的正确性，若失败可能影响配置。
- 日志添加可能引入不必要的输出噪声。

影响：

- 影响限于 CI 测试套件，确保 ROCm 平台翻译测试通过，对用户无直接影响。
- 提升开发者 CI 稳定性，减少 flakiness，有助于 ROCm 兼容性维护。

## 关联脉络

与本 PR 相关的历史 PR 包括：

- PR 38381："[ROCm][CI] Pin test_hybrid test to TRITON_ATTN on ROCm"，类似地固定 ROCm 测试中的注意力后端以提升稳定性。
- PR 37698："[ROCm][Bugfix] fix exception related to trust_remote_code for MiniMax-M2.1-MXFP4"，同为 ROCm 平台 bugfix，显示团队持续优化 ROCm 兼容性测试。

这些 PR 共同反映了 ROCm 生态中注意力后端配置的逐步完善。