

PR #46705 完整报告

vllm-project/vllm

Migrate Voxtral to mistral-common 1.11.5 audio API

合并时间: 2026-06-27 02:06

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46705>

执行摘要

- 一句话: 迁移 Voxtral 至 mistral-common 1.11.5 音频 API
- 推荐动作: 该 PR 属于常规维护, 展示了如何优雅地上游依赖 API 变化。对于维护多模态模型的开发人员有一定参考价值, 但非关键变更, 可按需阅读。

功能与动机

根据 PR body, 目的是「prevent deprecation warnings」, 即消除 mistral_common 旧版 API 的弃用警告, 并适配新版本音频 API。作者在 Issue 评论中提到上游存在回归问题, 需先在 mistral_common 修复后再更新版本。

实现拆解

1. 导入路径迁移: 将 `from mistral_common.audio import Audio` 替换为 `from mistral_common.tokens.tokenizers.audio import Audio`, 同时移除 `RawAudio` 的导入, 该符号在新版 API 中已废弃。涉及文件: `vllm/model_executor/models/voxtral.py`、`vllm/model_executor/models/voxtral_realtime.py`、`examples/generate/multimodal/audio_language_offline.py`、`tests/models/multimodal/generation/test_voxtral.py`、`tests/models/multimodal/generation/test_voxtral_realtime.py`。
2. 构造 API 替换:
 - 将 `AudioChunk(input_audio=RawAudio.from_audio(audio_item))` 改为 `AudioChunk.from_audio(audio_item)`——新版 `AudioChunk` 提供了直接从 `Audio` 对象构造的类方法。
 - 在 `get_generation_prompt` 中, 将 `TranscriptionRequest` 的 `audio` 参数从 `RawAudio.from_audio(audio)` 改为 `audio.to_base64(audio.format)`, 即直接以 Base64 编码传递音频。
3. 依赖版本升级: 在 `requirements/test/cuda.in`、`requirements/test/rocm.in` 等测试依赖配置中, 将 `mistral_common[image,audio]` 的最低版本从 `>=1.11.3` 提升至 `>=1.11.5`, 以确保使用经过回归修复的稳定版本。

关键文件:

- `vllm/model_executor/models/voxtral.py` (模块 多模态; 类别 source; 类型 data-contract; 符号 `get_dummy_processor_inputs`, `get_generation_prompt`,

get_speech_to_text_config) : Voxtral 模型主文件, 关键导入替换和构造 API 迁移, 影响核心逻辑

- vllm/model_executor/models/voxtral_realtime.py (模块 多模态; 类别 source; 类型 data-contract; 符号 get_generation_prompt) : Voxtral Realtime 模型主文件, 同样需要导入和 API 迁移
- examples/generate/multimodal/audio_language_offline.py (模块 示例; 类别 source; 类型 dependency-wiring; 符号 run_voxtral) : 官方示例, 展示 Voxtral 使用方式, 需要同步迁移
- tests/models/multimodal/generation/test_voxtral.py (模块 Voxtral 测试; 类别 test; 类型 test-coverage; 符号 _get_prompt) : Voxtral 核心测试, 确保迁移后功能正确
- tests/models/multimodal/generation/test_voxtral_realtime.py (模块 Voxtral 实时测试; 类别 test; 类型 test-coverage; 符号 from_file, test_voxtral_realtime_forward, test_voxtral_realtime_generator) : Voxtral Realtime 测试, 同步更新 API
- requirements/test/cuda.in (模块 依赖配置; 类别 config; 类型 test-coverage) : 测试依赖配置, 提升 mistral-common 最低版本

关键符号: VoxtralProcessingInfo.get_dummy_processor_inputs,
VoxtralForConditionalGeneration.get_generation_prompt,
VoxtralForConditionalGeneration.get_speech_to_text_config,
VoxtralRealtimeForConditionalGeneration.get_generation_prompt, run_voxtral
(example), _get_prompt (test)

关键源码片段

vllm/model_executor/models/voxtral_realtime.py

Voxtral Realtime 模型主文件, 同样需要导入和 API 迁移

```
# vllm/model_executor/models/voxtral_realtime.py (更改后)
# 新导入
from mistral_common.tokens.tokenizers.audio import Audio, AudioConfig

# 之前导入的 RawAudio 已移除
# ...

def get_generation_prompt(
    cls,
    stt_params: SpeechToTextParams,
) -> PromptType:
    audio = Audio(audio, int(stt_config.sample_rate), format="wav")
    req = TranscriptionRequest(
        model=model_config.model,
        audio=audio.to_base64(audio.format), # 之前: RawAudio.from_audio(audio)
        language=language,
        streaming=StreamingMode.OFFLINE,
    )
    # ...
```

评论区精华

作者在 Issue 评论中指出：由于 mistral-common 1.11.4 存在图像重排序的回归问题，导致测试失败，因此决定先在上游修复后发布 1.11.5 再更新此 PR。评论内容为：“Interesting there is a failure here. It is a regression test we lack in mistral-common regarding image re-ordering i'll first need to fix it in mistral-common, make a 1.11.5 and update here then :)”。此外无其他 reviewer 讨论，已由维护者 DarkLight1337 批准。

- mistral-common 回归问题导致测试失败 (other): 作者修复了 mistral-common 并发布 1.11.5，然后将 PR 更新至该版本。最终测试通过。

风险与影响

- 风险：主要风险在于依赖升级：如果 mistral-common 1.11.5 引入其他兼容性问题，可能影响 Voxtral 模型运行。但由于 PR 同时更新了测试和示例，且仅在 Voxtral 相关路径中修改，风险范围限于该模型。另外，audio.to_base64() 的替换可能改变音频传输格式，需确认服务端是否支持 Base64 编码；从代码看 TranscriptionRequest 预期此格式，风险较低。
- 影响：影响范围：仅影响使用 Voxtral 模型的用户，包括 Vixtral 和 Voxtral-Realtime 两个变体。影响程度：低——功能不变，仅内部 API 调用方式变化。用户无需修改代码，但需要升级 mistral-common 依赖至 $\geq 1.11.5$ 。CI 测试已覆盖核心路径，回归风险小。
- 风险标记：依赖升级，窄影响面

关联脉络

- 暂无明显关联 PR