

PR #45113 完整报告

vllm-project/vllm

Fix misleading error for audio duration limit rejection

合并时间: 2026-06-12 17:58

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45113>

执行摘要

- 一句话: 修复音频时长超限错误提示误导问题
- 推荐动作: 可合并, 属于小改进, 值得精读以了解错误传播的最佳实践。

功能与动机

当音频时长超过限制时, 用户收到 "Invalid or unsupported audio file" 的错误信息, 无法得知真正的限制原因和解决方法。PR body 明确指出: 让 `load_audio` 抛出的描述性 `ValueError` 直接传播, 并更新错误信息以提及 `VLLM_MAX_AUDIO_DECODE_DURATION_S` 环境变量。

实现拆解

1. 修改 `vllm/multimodal/media/audio.py` 中的错误信息: 在 `load_audio_pyav` 和 `load_audio_soundfile` 中, 将过时限制的 `ValueError` 信息从 "This limit prevents decompression-bomb attacks." 改为 "Set `VLLM_MAX_AUDIO_DECODE_DURATION_S` to increase this limit.", 让用户知道可以调整环境变量。
2. 修改 `vllm/entrypoints/speech_to_text/base/serving.py` 中的异常处理: 在 `_decode_and_chunk_speech` 方法中, 在 `try` 块之后增加一个 `except ValueError: raise` 子句, 确保来自 `load_audio` 的描述性 `ValueError` 能直接传播到上层, 而不是被后续的通用 `except Exception` 捕获并转换为模糊的 "Invalid or unsupported audio file"。

关键文件:

- `vllm/multimodal/media/audio.py` (模块 多模态; 类别 source; 类型 core-logic; 符号 `load_audio_pyav`, `load_audio_soundfile`): 核心文件: 修改了 `load_audio_pyav` 和 `load_audio_soundfile` 中的错误信息, 指引用户通过设置环境变量来解决时长限制问题。
- `vllm/entrypoints/speech_to_text/base/serving.py` (模块 前端入口; 类别 source; 类型 core-logic; 符号 `_decode_and_chunk_speech`): 关键入口: 修改了 `_decode_and_chunk_speech` 方法的异常处理, 确保 `load_audio` 抛出的 `ValueError` 不被通用异常捕获, 从而保留详细的错误信息。

关键符号: `load_audio_pyav`, `load_audio_soundfile`, `_decode_and_chunk_speech`

关键源码片段

`vllm/multimodal/media/audio.py`

核心文件：修改了 `load_audio_pyav` 和 `load_audio_soundfile` 中的错误信息，指引用户通过设置环境变量来解决时长限制问题。

```
# vllm/multimodal/media/audio.py
# 在 load_audio_pyav 和 load_audio_soundfile 中,
# 将原本的泛泛提示 "This limit prevents decompression-bomb attacks."
# 替换为具体的环境变量指引 "Set VLLM_MAX_AUDIO_DECODE_DURATION_S to increase this
limit."
# 让用户明确知道如何调整限制。
```

```
def load_audio_pyav(...):
    # ... 省略上下文
    if metadata_duration_s is not None and metadata_duration_s > max_duration_s:
        raise ValueError(
            f"Audio exceeds maximum allowed duration of "
            f"{max_duration_s}s (metadata reports "
            f"{metadata_duration_s:.1f}s). Set "
            f"VLLM_MAX_AUDIO_DECODE_DURATION_S to "
            f"increase this limit." # <-- 原先是 "This limit prevents decompression-bomb attacks."
        )
    # ... 省略后续
    if max_samples is not None and total_samples > max_samples:
        raise ValueError(
            f"Audio exceeds maximum allowed duration of "
            f"{max_duration_s}s (decoded {total_samples} "
            f"samples at {sr}Hz). Set "
            f"VLLM_MAX_AUDIO_DECODE_DURATION_S to "
            f"increase this limit." # <-- 同样修改
        )
```

```
def load_audio_soundfile(...):
    # ... 省略上下文
    if file_duration_s > max_duration_s:
        raise ValueError(
            f"Audio exceeds maximum allowed duration of "
            f"{max_duration_s}s (file contains "
            f"{file_duration_s:.1f}s at {native_sr}Hz). Set "
            f"VLLM_MAX_AUDIO_DECODE_DURATION_S to "
            f"increase this limit." # <-- 同样修改
        )
```

`vllm/entrypoints/speech_to_text/base/serving.py`

关键入口：修改了 `_decode_and_chunk_speech` 方法的异常处理，确保 `load_audio` 抛出的 `ValueError` 不被通用异常捕获，从而保留详细的错误信息。

```
# vllm/entrypoints/speech_to_text/base/serving.py
# 在 _decode_and_chunk_speech 方法中,
# 添加了 except ValueError: raise 子句,
# 使得 load_audio 抛出的描述性 ValueError 能直接传播到上层,
```

而不是被通用的 `except Exception` 捕获并替换为模糊的错误信息。

```
def _decode_and_chunk_speech(self, audio_data: bytes):
    try:
        with io.BytesIO(audio_data) as buf:
            y, sr = load_audio(
                buf,
                sr=self.asr_config.sample_rate,
                max_duration_s=self.max_audio_decode_duration_s,
            )
    except ValueError: # <-- 新增: 让 load_audio 的 ValueError 直接传播
        raise
    except Exception as exc:
        raise ValueError("Invalid or unsupported audio file.") from exc
    # 后续处理不变 ...
```

评论区精华

没有 review 评论，只有 approver 的批准和 mergify 的冲突提示。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。改动范围小（仅两个文件，共 11 行新增、6 行删除），主要是错误消息和异常处理链的调整。不会引入功能性回归。但需注意：若 `load_audio` 抛出非 `ValueError` 的其他异常（如 `OSError`），仍然会被捕获为 "Invalid or unsupported audio file"，这部分逻辑未变。
- 影响：对用户：当音频时长超过限制时，用户将看到明确指导如何增加限制的错误消息，而不是模糊的错误提示。对系统：无性能影响，异常处理逻辑更清晰。对团队：无显著影响。
- 风险标记：小范围变更，低风险

关联脉络

- PR #44612 [ASR] Optimize CPU preproc to get 2.5x RTFx via multi-threading: 同属 ASR 功能模块，修改了同一个 `serving.py` 文件中的相关方法。