

PR #45908 完整报告

vllm-project/vllm

fix(security): enforce audio decode duration limit in chat completions path

合并时间: 2026-06-17 19:07

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45908>

执行摘要

- 一句话: 修复音频解压炸弹防护绕过
- 推荐动作: 值得合并, 修复了一个安全遗漏。建议后续补充测试, 确保该限制不被后续重构绕过。

功能与动机

AudioMediaIO.load_bytes() 和 load_file() 在调用 load_audio() 时未传递 max_duration_s 参数, 导致通过 /v1/chat/completions 的 input_audio 内容部分加载的音频绕过了解压炸弹防护。已在 /v1/audio/transcriptions 路径中存在的 VLLM_MAX_AUDIO_DECODE_DURATION_S 限制需要同步应用到聊天补全路径。

实现拆解

1. 添加环境变量导入: 在 vllm/multimodal/media/audio.py 顶部新增 import vllm.envs as envs, 以访问 VLLM_MAX_AUDIO_DECODE_DURATION_S 配置项。
2. 修改 load_bytes 方法: 将 AudioMediaIO.load_bytes() 中对 load_audio() 的调用从 load_audio(BytesIO(data), sr=None) 改为 load_audio(BytesIO(data), sr=None, max_duration_s=envs.VLLM_MAX_AUDIO_DECODE_DURATION_S), 传入时长限制。
3. 修改 load_file 方法: 同理, 将 AudioMediaIO.load_file() 中的调用从 load_audio(filepath, sr=None) 改为 load_audio(filepath, sr=None, max_duration_s=envs.VLLM_MAX_AUDIO_DECODE_DURATION_S)。
4. 无测试或配置变更: 本次变更仅涉及源码文件, 未添加测试或修改配置。

关键文件:

- vllm/multimodal/media/audio.py (模块 多模态; 类别 source; 类型 dependency-wiring; 符号 load_bytes, load_file): 核心文件, 修复了音频加载时未传递时长限制的安全问题。

关键符号: load_bytes, load_file

关键源码片段

`vllm/multimodal/media/audio.py`

核心文件, 修复了音频加载时未传递时长限制的安全问题。

```
# vllm/multimodal/media/audio.py (partial)
```

```
import vllm.envs as envs # 新增：导入环境变量模块

class AudioMediaIO(MediaIO[tuple[npt.NDArray, float]]):
    # ...

    def load_bytes(self, data: bytes) -> tuple[npt.NDArray, float]:
        return load_audio(
            BytesIO(data),
            sr=None,
            # 新增：传递时长限制，防止解压炸弹
            max_duration_s=envs.VLLM_MAX_AUDIO_DECODE_DURATION_S,
        )

    def load_file(self, filepath: Path) -> tuple[npt.NDArray, float]:
        return load_audio(
            filepath,
            sr=None,
            # 新增：传递时长限制
            max_duration_s=envs.VLLM_MAX_AUDIO_DECODE_DURATION_S,
        )
```

评论区精华

PR 无 review 评论，由 DarkLight1337 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：该 PR 仅将已有的环境变量限制应用到两个调用点，不会改变现有行为（除非用户未设置该环境变量，此时默认值由 envs 模块处理）。但缺少测试覆盖，建议后续添加单元测试验证限制生效。
- 影响：影响范围：仅影响通过 /v1/chat/completions 路径上传音频的用户。增强了安全性，防止恶意长音频导致内存膨胀。对系统性能无负面影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR