

PR #49073 完整报告

vllm-project/vllm

[Bugfix][Frontend] Return transcription and translation verbose as float

合并时间: 2026-07-29 21:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49073>

执行摘要

- 一句话: 修复 verbose 响应 duration 字段类型为 float
- 推荐动作: 值得快速合并。变更简单明确, 已通过手动端到端验证。不建议精读代码, 但值得关注的决策是 reviewer 坚持“直接改定义, 不额外加测试”的务实风格。

功能与动机

修复 issue #49068 报告的问题: `verbose_json` 响应的 `duration` 字段以字符串形式返回, 违反 OpenAI 兼容 API 约定, 导致类型化客户端 (如 TypeScript SDK) 无法正确解析。

实现拆解

1. 修改协议模型字段类型: 在两个协议文件 `transcription/protocol.py` 和 `translation/protocol.py` 中, 将 `TranscriptionResponseVerbose.duration` 和 `TranslationResponseVerbose.duration` 的类型从 `str` 改为 `float`。
2. 移除构造时的字符串转换: 在 `base/serving.py` 的 `_create_speech_to_text` 方法中, 移除 `duration=str(duration_s)` 中的 `str()` 调用, 直接传递浮点数 `duration_s`。
3. 移除冗余测试: 最初提交中新增了测试文件 `test_verbose_response_protocol.py`, 但 reviewer 认为用处不大, 提议直接改定义, 最终删除该测试。

关键文件:

- `vllm/entrypoints/speech_to_text/base/serving.py` (模块 前端服务; 类别 `source`; 类型 `core-logic`; 符号 `_create_speech_to_text`): 核心逻辑变更: 移除 `duration` 的 `str()` 转换, 确保直接传入浮点数。
- `vllm/entrypoints/speech_to_text/transcription/protocol.py` (模块 前端协议; 类别 `source`; 类型 `core-logic`; 符号 `TranscriptionResponseVerbose`): 类型定义变更: 将 `duration` 字段类型从 `str` 改为 `float`, 影响 JSON 序列化行为。
- `vllm/entrypoints/speech_to_text/translation/protocol.py` (模块 前端协议; 类别 `source`; 类型 `core-logic`; 符号 `TranslationResponseVerbose`): 与 `transcription` 协议镜像修改: `duration` 字段类型从 `str` 改为 `float`。

关键符号: `_create_speech_to_text`

关键源码片段

vllm/entrypoints/speech_to_text/base/serving.py

核心逻辑变更：移除 duration 的 str() 转换，确保直接传入浮点数。

```
# vllm/entrypoints/speech_to_text/base/serving.py
# 在 _create_speech_to_text 方法中，构造 verbose 响应时
# 移除 str() 包装，使 JSON 序列化输出 number 而非 string
if request.response_format != "verbose_json":
    final_response = cast(
        T, TranscriptionResponse(text=text, usage=usage)
    )
else:
    final_response = cast(
        V,
        TranscriptionResponseVerbose(
            text=text,
            language=request.language,
            duration=duration_s, # 原为 str(duration_s)
            segments=total_segments,
        ),
    )
```

vllm/entrypoints/speech_to_text/transcription/protocol.py

类型定义变更：将 duration 字段类型从 str 改为 float，影响 JSON 序列化行为。

```
# vllm/entrypoints/speech_to_text/transcription/protocol.py
class TranscriptionResponseVerbose(OpenAIBaseModel):
    duration: float # 原为 str，改为 float 以符合 OpenAI 规范
    """The duration of the input audio."""
    language: str
    text: str
    segments: list[TranscriptionSegment] | None = None
    words: list[TranscriptionWord] | None = None
```

vllm/entrypoints/speech_to_text/translation/protocol.py

与 transcription 协议镜像修改：duration 字段类型从 str 改为 float。

```
# vllm/entrypoints/speech_to_text/translation/protocol.py
class TranslationResponseVerbose(OpenAIBaseModel):
    duration: float # 原为 str，改为 float 以符合 OpenAI 规范
    """The duration of the input audio."""
    language: str
    text: str
    segments: list[TranslationSegment] | None = None
    words: list[TranslationWord] | None = None
```

评论区精华

核心讨论围绕新增的测试文件展开。Reviewer DarkLight1337 认为测试“没那么有用，直接改定义就行”，作者 wskr00 同意并删除了文件。其他评论仅涉及简单的流程确认。

- 测试文件必要性 (testing): 确认移除测试文件, 仅做协议定义修改。

风险与影响

- 风险: 风险极低。仅修改了字段类型和一处类型转换, 改动范围小 (共 4 行), 且改动的逻辑位于非关键路径。不涉及性能、安全或兼容性问题。需确保下游消费者未依赖 duration 为字符串类型 (理论上不符合规范, 可能性低)。
- 影响: 影响范围仅限于使用 response_format=verbose_json 的 /v1/audio/transcriptions 和 /v1/audio/translations 端点。修复后 duration 字段将以 JSON number 返回, 提升了 API 兼容性和客户端易用性。无负面兼容性影响。
- 风险标记: 低风险

关联脉络

- PR #49068 [Bug]: verbose_json returns duration as a string instead of a number: 本 PR 直接修复该 issue 报告的问题。