PR #39116 完整报告

vllm-project/vllm

[ASR] Fix spacing bw chunks in multi chunk audio transcription

合并时间：2026-04-10 03:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39116

执行摘要

该 PR 修复了 vLLM 中自动语音识别（ASR）在多块音频转录时块间缺少空格的问题，通过引入语言特定分隔符逻辑，在流式和非流式路径中正确插入空格，显著提升 Cohere 和 Qwen3 ASR 模型的输出质量。新增测试确保修复覆盖各种场景。

功能与动机

当前 ASR 转录在处理多块音频时，连续块的文本输出间无空格分隔，导致文本粘连（如 "Hello, thisis vllm"）。此问题影响 Cohere ASR 和 Qwen3 ASR，而 Whisper 模型因总是预加空格未被发现。修复旨在确保输出文本格式正确，提升用户体验。

实现拆解

- 核心辅助函数：在 speech_to_text.py 中新增 asr_inter_chunk_separator 函数，根据语言代码返回空格或空字符串（默认日语和中文无空格）。
- 非流式路径：在 _create_speech_to_text 方法中计算分隔符，使用 separator.join(text_parts) 合并文本块。
- 流式路径：更新 _speech_to_text_stream_generator、transcription_stream_generator 和 translation_stream_generator 方法，传递分隔符参数，并确保仅在非第一个块前添加分隔符。
- 协议扩展：在 SupportsTranscription 协议中添加 no_space_languages 类变量，提供默认值，并在 Cohere ASR 模型中显式定义。
- 测试覆盖：新增测试文件 test_transcription_inter_chunk_spacing.py，验证分隔符逻辑和流式 / 非流式行为。

评论区精华

review 中，gemini-code-assist[bot] 指出关键问题：

"The translation_stream_generator method ... has not been updated to accept the new separator argument. This will cause a `TypeError"

"Accessing self.model_cls.no_space_languages directly may raise an AttributeError ... Using getattr is safer"

"In streaming mode, the separator is prepended to every chunk, including the first one. This results in an unwanted leading space"

作者 ekagra-ranjan 回应认为所有转录模型应继承 SupportsTranscription 协议，因此无需额外安全措施。最终 PR 被批准，提交历史显示流式前导空格问题已修复。

风险与影响

风险：

- 属性访问风险：如果模型未定义 no_space_languages，直接访问可能引发 AttributeError，尽管协议有默认值。
- 流式逻辑错误：需确保分隔符仅在非第一个块添加，否则导致前导空格，提交显示已修复。
- 协议兼容性：新增协议变量需现有模型适配，但默认值应覆盖常见情况。

影响：

- 用户：ASR 输出质量提升，特别是英语等需要空格的语言。
- 系统：修改限于前端入口点，性能影响轻微。
- 团队：新增测试增强可靠性，协议变更需未来模型实现遵循。

关联脉络

与近期 PR 关联较弱，但 PR 38538（Nemotron-Nano-VL 音频处理）和 PR 39409（多模态错误信息）同属前端和多模态领域，显示团队在完善 ASR 和相关功能。此 PR 聚焦于 ASR 转录的细节修复，是整体语音处理改进的一部分。