PR #39845 完整报告

vllm-project/vllm

[Doc] Add Realtime Transcription section to supported_models.md

合并时间：2026-04-18 11:26

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39845

执行摘要

- 一句话：在支持模型文档中添加实时转录模型章节并修复 API 文档中的错误链接。
- 推荐动作：此 PR 是一个简单的文档完善，无需深入技术分析。对于工程师，如果正在处理实时转录相关功能或需要参考模型支持列表，可以快速浏览 supported_models.md 中的新章节以获取关键信息。对于技术管理者，此变更展示了文档维护的重要性，但无需投入额外审查精力。

功能与动机

根据 PR 描述，当前 supported_models.md 文档没有列出实时流式架构，尽管 VoxtralRealtimeGeneration 和 Qwen3ASRRealtimeGeneration 已在 registry.py 中注册并通过 /v1/realtime 端点提供服务。同时，openai_compatible_server.md 中的 Realtime API 部分错误地链接到了批量转录章节（#transcription），而不是实时转录章节。此 PR 旨在填补文档空白并修正错误链接，确保用户能准确找到实时转录模型的使用信息。

实现拆解

1. 在支持模型文档中添加实时转录章节：修改 docs/models/supported_odels.md，在现有“Transcription”部分之后新增“Realtime Transcription”章节。该章节以表格形式列出支持的架构（VoxtralRealtimeGeneration 和 Qwen3ASRRealtimeGeneration）、对应的 HuggingFace 模型示例，并添加了使用注意事项（如 VoxtralRealtimeGeneration 需要 --tokenizer-mode mistral，Qwen3ASRRealtimeGeneration 需要 --hf-overrides）。
2. 修正 API 文档中的交叉引用：修改 docs/serving/openai_compatible_server.md，将 Realtime API 部分中“Only applicable to...”的链接从 #transcription 更正为 #realtime-transcription，确保指向新添加的实时转录章节。
3. 验证与提交：作者在 PR 描述中说明了验证步骤，包括检查模型在 registry.py 中的注册、确认 --hf-overrides 要求、验证链接解析正确性以及 Markdown 格式渲染。提交历史显示两个提交：第一个提交添加了文档变更，第二个提交是合并主分支的更新。

关键文件：

- docs/models/supported_models.md（模块 模型文档；类别 docs；类型 documentation）：这是核心变更文件，新增了实时转录模型的完整文档章节，包括架构列表、模型示例和使用注意事项，直接解决了 PR 的主要动机。

- docs/serving/openai_compatible_server.md（模块 服务文档；类别 docs；类型 documentation）：次要变更文件，修正了 Realtime API 部分中的错误交叉引用，确保链接指向新添加的实时转录章节，提升了文档内部一致性。

关键符号：未识别

评论区精华

review 中讨论较少，主要确认了变更的正确性。

- gemini-code-assist[bot]评论："此拉取请求在支持模型文档中添加了'实时转录'部分，特别强调了 VoxtralRealtimeGeneration 和 Qwen3ASRRealtimeGeneration。它还更新了实时 API 文档以链接到这个新部分。我没有反馈提供。"
- DarkLight1337评论："抱歉错过了这个，LGTM"。 讨论中没有出现争议点，变更被直接批准。
- 文档变更确认 (documentation): 变更被批准，认为 LGTM（Looks Good To Me）。

风险与影响

- 风险：此 PR 为纯文档更新，不涉及任何代码、配置或运行时逻辑的修改，因此不存在技术风险（如回归、性能、安全或兼容性问题）。唯一潜在风险是文档内容准确性，但作者已在 PR 描述中说明已验证模型注册和链接解析，且 reviewer 未提出异议，风险极低。
- 影响：影响范围：仅影响文档，特别是 supported_models.md 和 openai_compatible_server.md 两个文件。 影响程度：
- 对用户：正面影响，用户现在可以在官方文档中找到实时转录模型的支持列表和使用说明，避免了混淆和错误链接，提升了文档的完整性和可用性。
- 对系统：无影响，不改变任何系统行为或性能。
- 对团队：轻微影响，维护了文档与代码实现的一致性，减少了用户支持负担。
- 风险标记：文档准确性风险

关联脉络

- PR #38405 [Frontend] Add multimodal support to /inference/v1/generate endpoint: 同属文档更新类别，都涉及完善 API 端点（如 /v1/realtime）的文档支持，反映了 vLLM 在多模态和实时功能上的文档演进。
- PR #39291 feat: Add LoRA support for Gemma4ForConditionalGeneration: 都涉及模型支持文档的更新，PR#39291 为 Gemma4 模型添加 LoRA 支持说明，而本 PR 为实时转录模型添加支持说明，共同丰富了 supported_models.md 的内容。