PR #36679 完整报告

vllm-project/vllm

[Bugfix] stream failure when model name not in audio endpoints

合并时间：2026-04-13 22:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/36679

执行摘要

修复了音频转录和翻译端点在流式推理时，因请求未提供模型名称而导致的 Pydantic 验证错误（400 BadRequest）。通过为缺失的模型名称设置默认值（self.models.model_name()），使该端点行为与其他 API 端点保持一致，确保流式推理正常工作。

功能与动机

问题背景：在使用 /v1/audio/transcriptions 或 /v1/audio/translations 端点进行流式推理（stream=true）时，如果请求中未包含 model 参数，会返回 Pydantic 验证错误：

```
{"error": {"message": "1 validation error for TranscriptionStreamResponse\nmodel\n Input should be a valid string [type=string_type, input_value=None, input_type=NoneType]", ...}}
```

而其他端点（如聊天完成）在类似情况下可以正常工作。PR body 中提供了完整的 curl 命令示例，展示了修复前后的对比。

实现拆解

仅修改了一个文件：vllm/entrypoints/openai/speech_to_text/speech_to_text.py。

在 _create_speech_to_text 函数中，添加了 3 行代码：

```
if not request.model:
    request.model = self.models.model_name()
```

位置在错误检查之后、引擎状态检查之前。这确保了当 request.model 为 None 时，使用 self.models.model_name() 提供的默认模型名称，从而通过后续的 Pydantic 验证。

评论区精华

review 讨论较少，主要来自 gemini-code-assist[bot] 的自动评论：

> "This pull request addresses a bug in the speech-to-text streaming endpoint where requests without a model name would fail. The change introduces logic to assign a default model name if one is not provided, aligning the endpoint's behavior with other parts of the API. The implementation is straightforward and correctly resolves the issue."

其他 reviewer（DarkLight1337 和 NickLucche）仅批准，无额外评论。

风险与影响

技术风险：

- 低风险：变更仅涉及前端请求验证逻辑，不触及核心推理路径。
- 依赖 self.models.model_name() 返回有效的模型名称，需确保该属性在上下文中已正确初始化。
- 无测试文件变更，但 PR 描述中已通过 curl 命令验证修复。

影响范围：

- 用户：修复了音频流式推理的可用性问题，提升用户体验。
- 系统：仅影响前端 API 端点，对性能、安全性无显著影响。
- 团队：代码变更极小，易于理解和维护。

关联脉络

与近期历史 PR 的关联：

- PR #37727：修复 Responses API 中参数泄漏问题，同属 frontend 模块的 bugfix。
- PR #38827：为 rerank 请求添加新参数，同属 frontend 模块的 API 增强。

这表明团队持续在完善前端 API 的健壮性和功能一致性。本 PR 是这一趋势的一部分，确保音频端点与其他端点行为对齐。