

PR #46197 完整报告

vllm-project/vllm

[Docs] Add Qwen3 forced alignment online example

合并时间: 2026-06-23 23:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46197>

执行摘要

- 一句话: 新增 Qwen3 强制对齐在线 /pooling 示例脚本
- 推荐动作: 建议用户浏览此 PR 作为在线 /pooling 请求的参考实现, 特别是如何结合音频与参考文本进行 token 分类任务。

功能与动机

来自 Issue #38805, 用户指出 vLLM 仅提供离线强制对齐示例, 缺少在线 /pooling 的用法说明。此 PR 旨在为社区提供一个可直接运行的脚本, 降低在线推理的上手门槛。

实现拆解

脚本按以下步骤实现:

1. 参数解析: 通过 `argparse` 解析服务器地址、模型名称、音频路径和参考词列表, 默认使用 5 秒静音 WAV。
2. Prompt 构建: `build_prompt` 将单词列表拼接成包含 `<audio_startl>` 和 `<timestamp>` 标记的字符串。
3. 音频编码: `encode_audio_data_uri` 或 `encode_silent_wav_data_uri` 将本地音频或生成静音波编码为 data URI 格式。
4. Payload 构建: `build_payload` 构造 /pooling 请求体, 包含模型名、消息列表、任务类型 `token_classify` 和自定义 chat template。
5. 请求发送与解析: `post_http_request` 发送 HTTP POST 请求, `parse_response` 验证响应状态并返回 JSON。
6. 时间戳解码: `load_timestamp_config` 从 HuggingFace 配置读取 `timestamp_token_id` 和 `timestamp_segment_time`, 在 main 中结合预测的 logits 计算每个单词的起始 / 结束时间并打印。

关键文件:

- `examples/pooling/token_classify/forced_alignment_online.py` (模块 示例; 类别 `source`; 类型 `example`; 符号 `build_prompt`, `encode_audio_data_uri`, `encode_silent_wav_data_uri`, `build_payload`): 新增的在线强制对齐示例脚本, 完整展示了 /pooling 请求的构造与响应解析过程。

关键符号: build_prompt, encode_audio_data_uri, encode_silent_wav_data_uri, build_payload, post_http_request, parse_response, load_timestamp_config, parse_args, main

评论区精华

审阅者 yewentao256 要求作者贴出脚本的完整运行输出，以验证结果格式。作者在 PR body 中提供了默认静音 WAV 的输出示例，输出格式为 `Hello 0.000s - 0.160s world 0.160s - 1.920s`。随后 yewentao256 确认并批准 PR，认为这解决了 Issue #38805。

- 运行输出确认 (question): 作者在 PR body 中提供了默认静音 WAV 的输出示例，显示格式为 `Hello 0.000s - 0.160s world 0.160s - 1.920s`。审阅者确认后批准 PR。
- Issue 关联确认 (other): 审阅者 yewentao256 批准 PR 时明确表示此 PR 关闭了 Issue #38805。

风险与影响

- 风险：风险极低，仅新增示例脚本，未修改任何核心代码。但用户需注意依赖库（如 hf_hub_download、pybase64）的网络可用性，以及模型的配置键名可能变化。
- 影响：对希望在线使用 Qwen3 强制对齐的用户有直接帮助，提供了可复用的模板代码。对系统无影响。
- 风险标记：无代码库变更，依赖外部网络和 HuggingFace Hub

关联脉络

- PR #38805 [Usage]: Does vllm support online infer for qwen3_asr_forced_aligner now? I only found offline example: 触发此 PR 的 Issue，用户请求在线推理示例。