

PR #46793 完整报告

vllm-project/vllm

[Frontend] Support bad_words in the /v1/completions endpoint

合并时间: 2026-07-08 17:51

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46793>

执行摘要

- 一句话: 为 /v1/completions 端点添加 bad_words 支持
- 推荐动作: 值得合并, 作为小而清晰的功能对等性补丁。建议阅读 CompletionRequest 模型中新增字段的位置以及 to_sampling_params() 中对应的传参行, 以理解 vLLM 中如何将前端请求参数映射到内部采样参数。

功能与动机

聊天补全端点 (/v1/chat/completions) 已支持 bad_words 并转发到 SamplingParams, 但旧版补全端点 (/v1/completions) 却缺少该功能, 尽管 SamplingParams 已完全支持 bad_words。这是一个功能对等性缺口: 使用补全端点的用户无法抑制特定词汇。PR 描述中明确说明了这是功能对齐需求, 且经搜索确认为首次添加。

实现拆解

1. 在 CompletionRequest 模型中新增字段: 在 vllm/entrypoints/openai/completion/protocol.py 的 CompletionRequest 类中的采样参数区 (# --8<-- [start:completion-sampling-params]) 添加 bad_words: list[str] = Field(default_factory=list), 默认值为空列表, 与聊天端点行为一致。
2. 在 to_sampling_params() 方法中透传参数: 在同一个文件中, 在 SamplingParams.from_optional() 调用中添加 bad_words=self.bad_words, 确保该字段被正确转发到底层的采样参数对象。
3. 添加单元测试: 在 tests/entrypoints/openai/completion/test_completion.py 中新增两个测试函数:
 - test_completion_request_bad_words_to_sampling_params: 验证 bad_words 被正确转发到 SamplingParams。
 - test_completion_request_bad_words_default_empty: 验证未指定 bad_words 时默认值为空列表。同时更新了文件导入, 添加了对 CompletionRequest 和 SamplingParams 的引用。

关键文件:

- vllm/entrypoints/openai/completion/protocol.py (模块 请求协议; 类别 source; 类型 core-logic): 核心变更文件: 在 CompletionRequest 模型中新增了 bad_words 字段, 并在 to_sampling_params() 方法中将其传递给 SamplingParams。

- tests/entrypoints/openai/completion/test_completion.py (模块测试; 类别 test; 类型 test-coverage; 符号 test_completion_request_bad_words_to_sampling_params, test_completion_request_bad_words_default_empty) : 测试配套文件: 新增两个单元测试覆盖了 bad_words 的默认值和转发逻辑, 确保行为与聊天端点一致。

关键符号: to_sampling_params

关键源码片段

vllm/entrypoints/openai/completion/protocol.py

核心变更文件: 在 `CompletionRequest` 模型中新增了 `bad_words` 字段, 并在 `to_sampling_params()` 方法中将其传递给 `SamplingParams`。

```
# vllm/entrypoints/openai/completion/protocol.py

class CompletionRequest(OpenAIBaseModel):
    ...
    # 在采样参数区域新增 bad_words 字段, 默认空列表, 与聊天端点对齐
    bad_words: list[str] = Field(default_factory=list)
    # --8<-- [end:completion-sampling-params]

    ...

    def to_sampling_params(...):
        ...
        return SamplingParams.from_optional(
            ...
            allowed_token_ids=self.allowed_token_ids,
            bad_words=self.bad_words, # 新添加的透传参数
            extra_args=extra_args or None,
            ...
        )
```

评论区精华

评论较少, 主要来自审核者 [DarkLight1337](#) 的快速批准 (仅回复 "Thanks")。claude[bot] 因为 PR 来自 fork 而跳过了自动审查。无 review 评论线程, 表明变更清晰、争议小。

- 暂无高价值评论线程

风险与影响

- 风险: 变更风险较低: 仅涉及前端的请求模型和参数透传, 不影响核心推理路径或后端逻辑。bad_words 字段使用 Pydantic 的 default_factory 确保默认行为安全。潜在风险在于用户可能误用或传递过长的列表, 但该风险与已有聊天端点一致, 且非此 PR 引入。
- 影响:
 - 用户影响: 使 /v1/completions 用户能够通过 bad_words 字段抑制指定词汇, 提升功能对等性。

- 系统影响：几乎无影响，仅增加了一个可选字段的解析和透传，不改变现有行为。
- 团队影响：降低因功能缺口导致的用户困惑或工单；代码改动量小，易于维护。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR