# PR #40532 完整报告

vllm-project/vllm

[Doc] Add missing API endpoints to security documentation

合并时间：2026-04-29 05:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40532

## 执行摘要

- 一句话：补全安全文档缺失的 API 端点列表
- 推荐动作：建议合并，无代码风险。该 PR 是安全文档的及时补充，值得管理员和部署者阅读以了解最新 endpoint 列表和认证要求。

## 功能与动机

安全文档中缺少多个实际存在的 API 端点，导致用户对哪些端点需要身份认证、哪些操作控制端点可被利用的理解不完整。PR body 明确指出："Several endpoints were missing from the API key authentication limitations section: RLHF weight-manipulation endpoints, generative scoring, abort_requests, and various /v1 sub-paths."

## 实现拆解

1. 更新 API 密钥认证章节：在 docs/usage/security.md 中添加了 /v1/chat/completions/batch、/v1/chat/completions/render、/v1/completions/render、/v1/messages/count_tokens、/v1/responses/{response_id}、/v1/responses/{response_id}/cancel、/v1/load_lora_adapter、/v1/unload_lora_adapter 等端点，并标注 LoRA 管理端点的开启条件和安全提示。
2. 更新无需认证端点章节：添加 /generative_scoring、/is_paused、/is_scaling_elastic_ep、/init_weight_transfer_engine、/update_weights、/get_world_size、/abort_requests 等端点，并注明部分端点的前置条件（如 --tokens-only）。
3. 修正操作控制端点描述：将 Operational control endpoints (always enabled) 改为 (only when "generate" task is supported)，反映实际的行为依赖。
4. 更新安全隐患描述：在安全影响部分补充了新增端点可能带来的攻击面（如权重操作、权重传输初始化和弹性缩放状态查询），并拆分 LoRA 端点作为单独警告段落（第二次提交新增）。

关键文件：

- docs/usage/security.md（模块 文档；类别 docs；类型 documentation）：唯一变更文件，系统性地补全了安全文档中认证和非认证 API 端点列表，并修正了操作控制端点的启用条件描述。

关键符号：未识别

## 评论区精华

审核来自 Claude 和 Gemini 的机器人自动评论，均未提出实质反馈；仓库维护者 sfeng33 直接批准。未发现人工评审讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：该 PR 仅涉及文档修改，无代码变更，无回归、性能或安全风险。但若文档描述与实际行为不符（例如新的 endpoint 是否存在或条件描述有误），可能误导管理员的安全配置。
- 影响：直接影响安全文档的完整性，帮助管理员正确配置 API 密钥认证和暴露范围；无用户功能变化或系统影响。
- 风险标记：暂无

## 关联脉络

- 暂无明显关联 PR