漏洞名称：

Xinference 存在未授权访问漏洞

厂商：

Xinference属于杭州未来速度科技有限公司

## Xinference 介绍

Xorbits Inference (Xinference) 是一个开源平台，用于简化各种 AI 模型的运行和集成。借助 Xinference，您可以使用任何开源 LLM、嵌入模型和多模态模型在云端或本地环境中运行推理，并创建强大的 AI 应用。

GitHub：<https://github.com/xorbitsai/inference>

官方文档：<https://inference.readthedocs.io/zh-cn/latest/index.html>

## 版本：

Xinference开源框架全版本（未设置访问认证的情况下）

## 漏洞描述：

Xinference默认未设置身份验证和访问控制功能。当Xinference拉起本地服务，且设置host为0.0.0.0时，攻击者可以远程条件下调用Xinference的API接口以及访问Web页面，执行包括但不限于敏感模型资产窃取、模型计算资源滥用、模型恶意删除和拒绝服务等恶意操作。

未设置身份验证和访问控制功能且暴露在公共互联网上的Xinference易受此影响

## 漏洞代码分析

1. 本地部署Xinference并启用

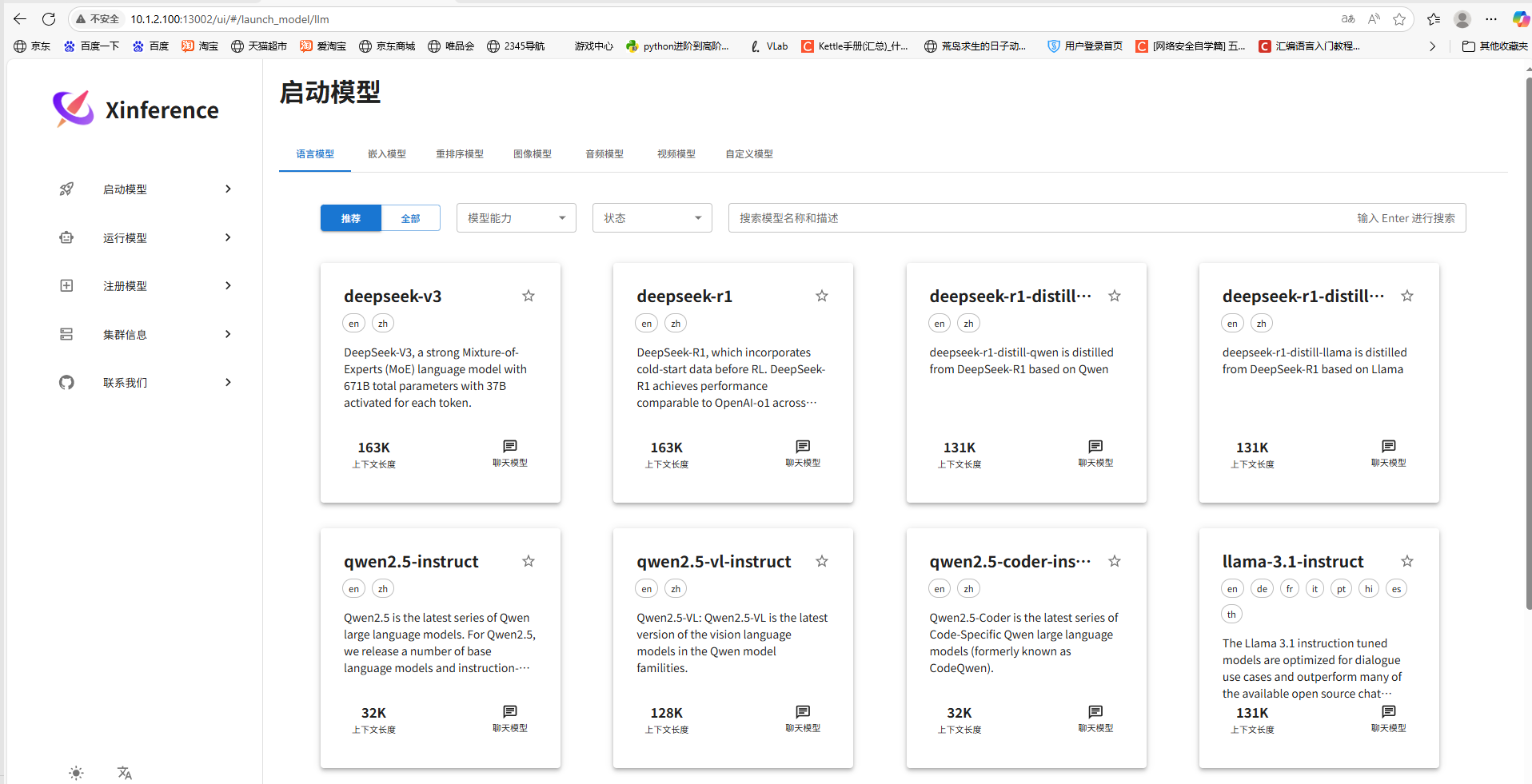
pip install "xinference[all]"

1. 默认配置情况下，本地运行Xinference的命令如下：

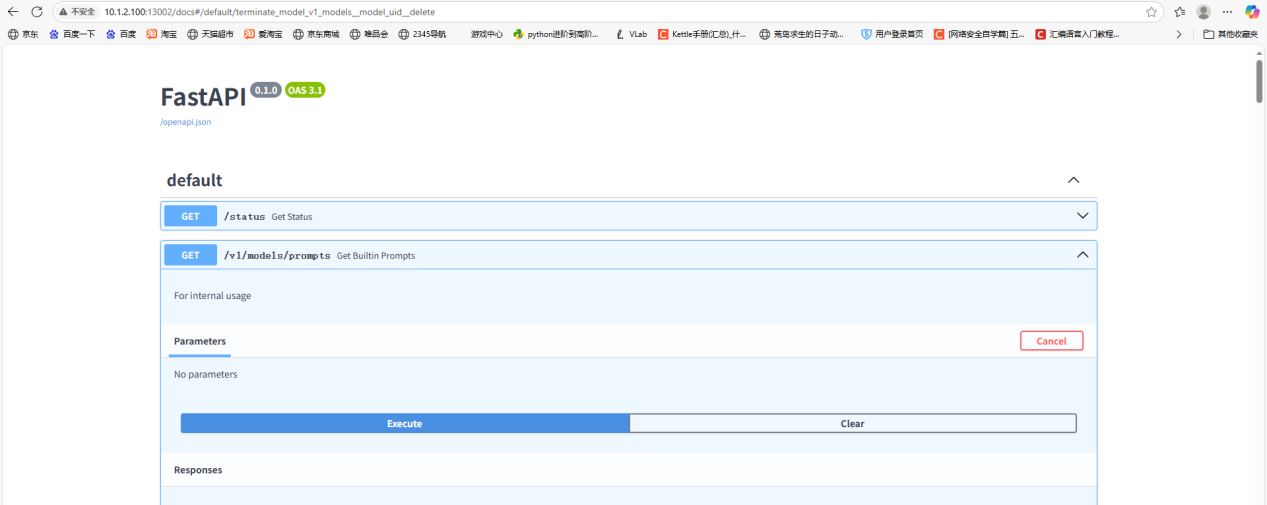
xinference-local --host 0.0.0.0 --port 9997

1. 针对公网暴露的Xinference，任意用户均可访问Xinference web及fastapi

访问web的路由地址：/ui/#/launch\_model/llm



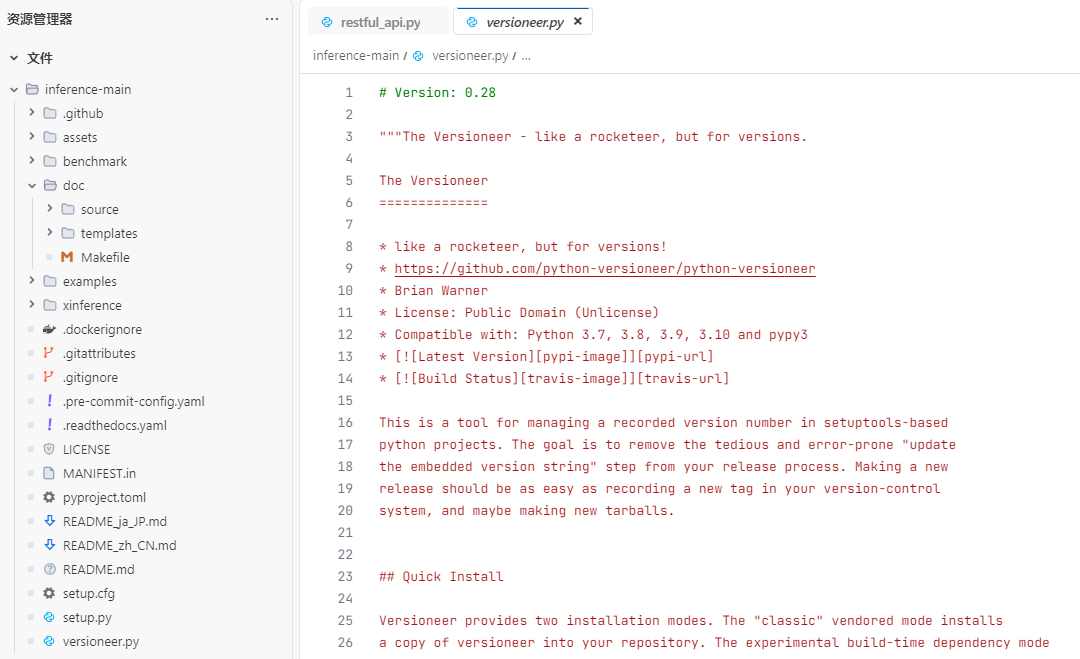
访问fastapi的路由地址：/docs



此时需要说明一下，xinference默认配置下，任何用户均可通过访问web页面的功能及fastapi接口，实现针对模型的拉取、查看、删除、篡改等功能。

源代码分析

本例中下载的xinference源码版本为0.28，



进入xinference/api/restful\_api.py源码文件中，可以看到fastapi的相关路由和接口



分析源码，可知：除/v1/models/prompts、/v1/models/vllm-supported

、/v1/cluster/info等路由外，其他路由均存在权限验证的身份依赖。



再次分析路由的权限验证身份依赖，源码如下。

此处对源码进行解读：如果想要执行/v1/models/{model\_uid} 路由，那么必须通过dependencies。

**在dependencies中，**

**如果self.is\_authenticated()的值为Ture，即系统启用认证的情况下，为 API 路由添加一个权限检查依赖项，要求访问该路由的用户必须具有 models:stop 权限。**

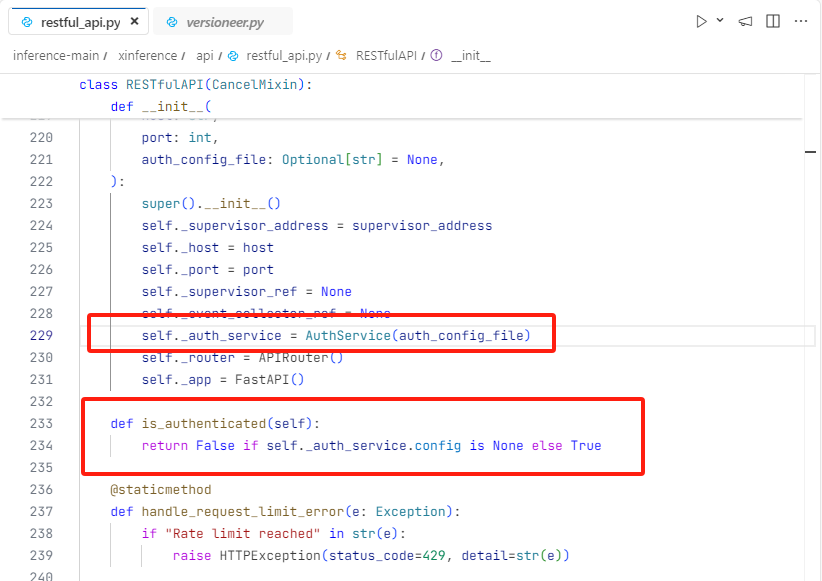
**如果系统未启用认证，则不进行权限检查。**



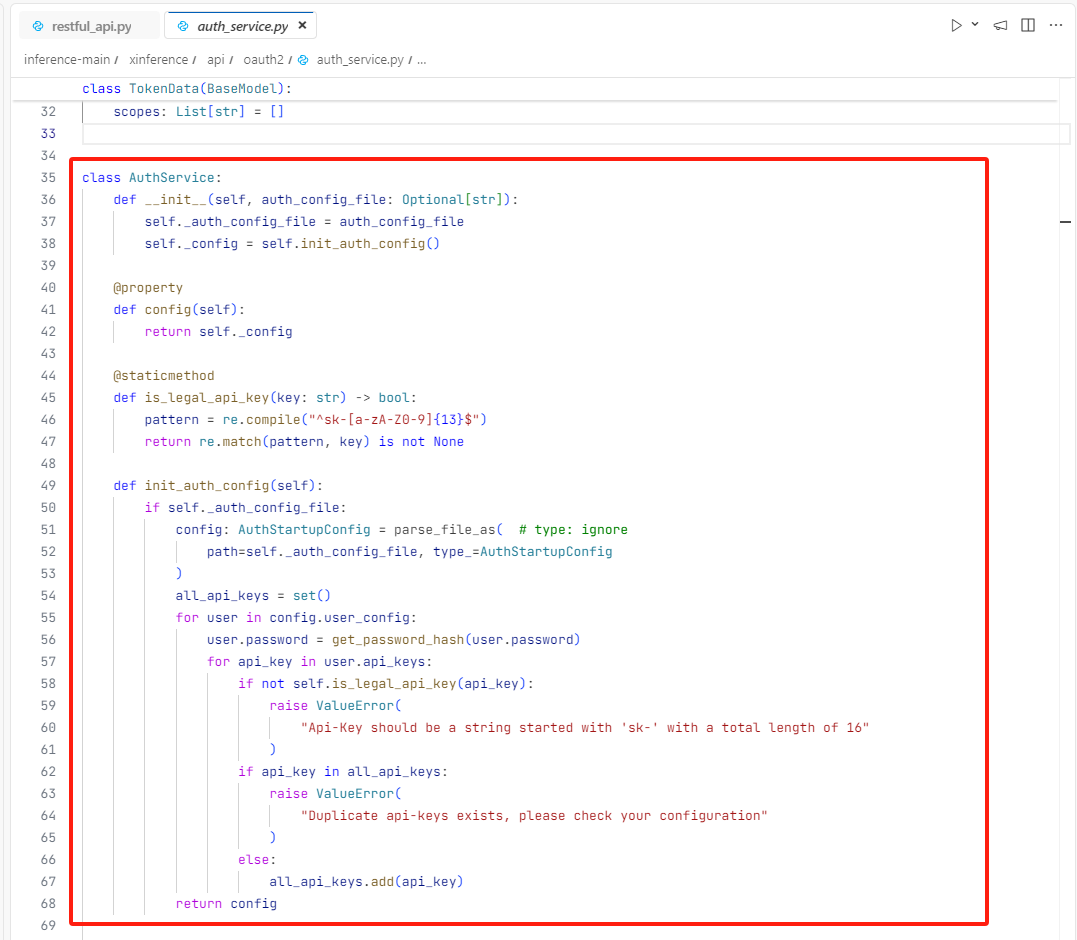
继续追踪self.is\_authenticated() 方法，代码如下：

发现self.is\_authenticated()的值与self.\_auth\_service.config 有关，若self.\_auth\_service.config为空，则返回False。

而通过self.\_auth\_service 可以追踪到AuthService(auth\_config\_file)方法。



AuthService方法存在与xinference/api/oauth2中，相关源码如下：



通过上述源码，我们进一步分析，可以得到结论如下：

\_auth\_service.config的值就是AuthService类中self.\_config的值，而self.\_config 的值由self.init\_auth\_config()返回。

而self.init\_auth\_config()返回的值，由auth\_config\_file来确定。

重点：如果auth\_config\_file为空，则self.init\_auth\_config()返回的值也为空。

在RESTfulAPI路由中，auth\_config\_file的值是多少呢？进一步查看，发现默认为空。



因此可以得出结论如下：

Xinference默认情况下auth\_config\_file 文件为空，导致权限默认为未设置。

Xinference默认配置下，攻击者可以未授权访问任意FastAPI路由，实现模型调用、模型删除等操作。

## 漏洞复现

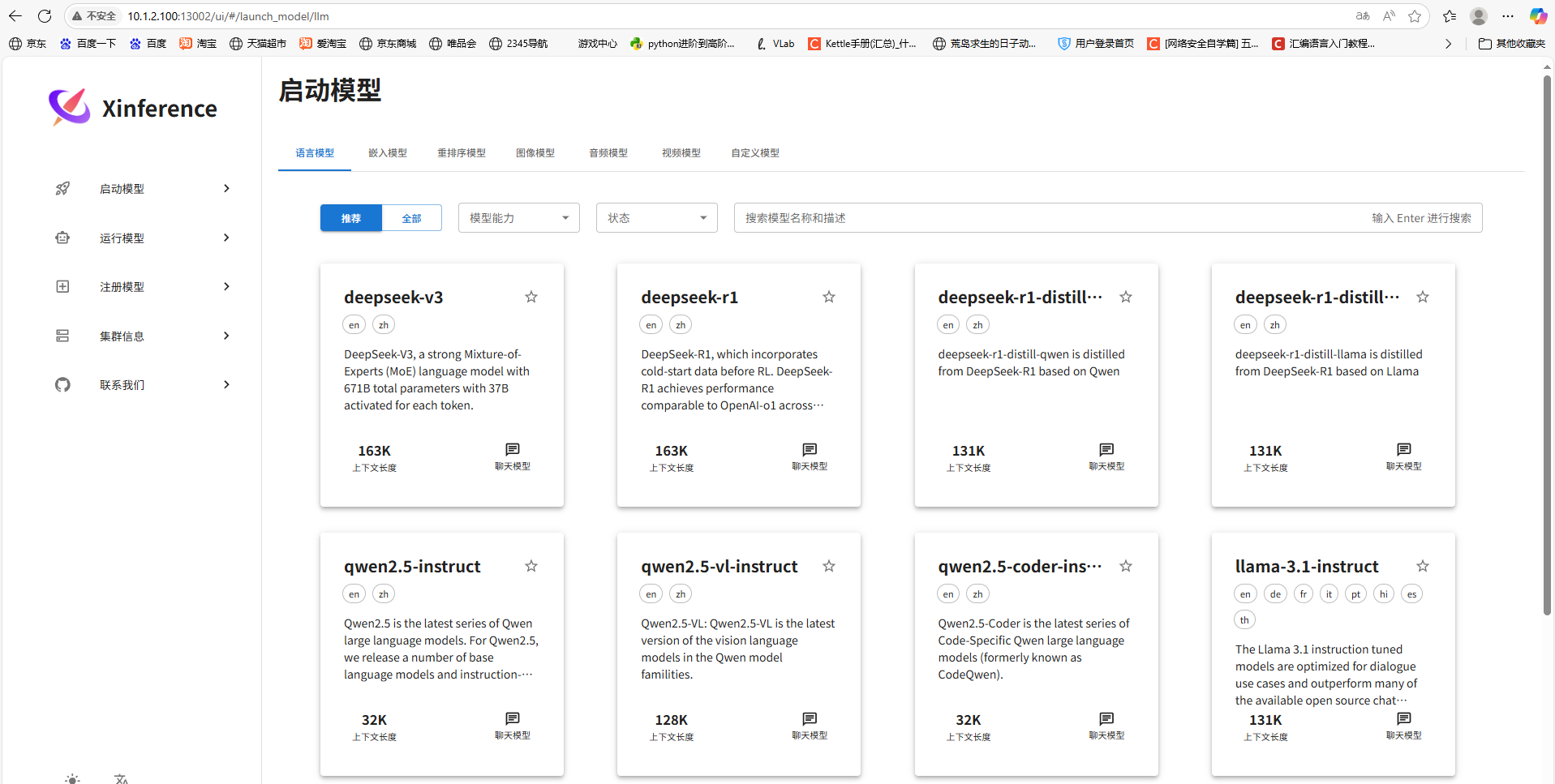
1. 本地部署Xinference并启用

pip install "xinference[all]"

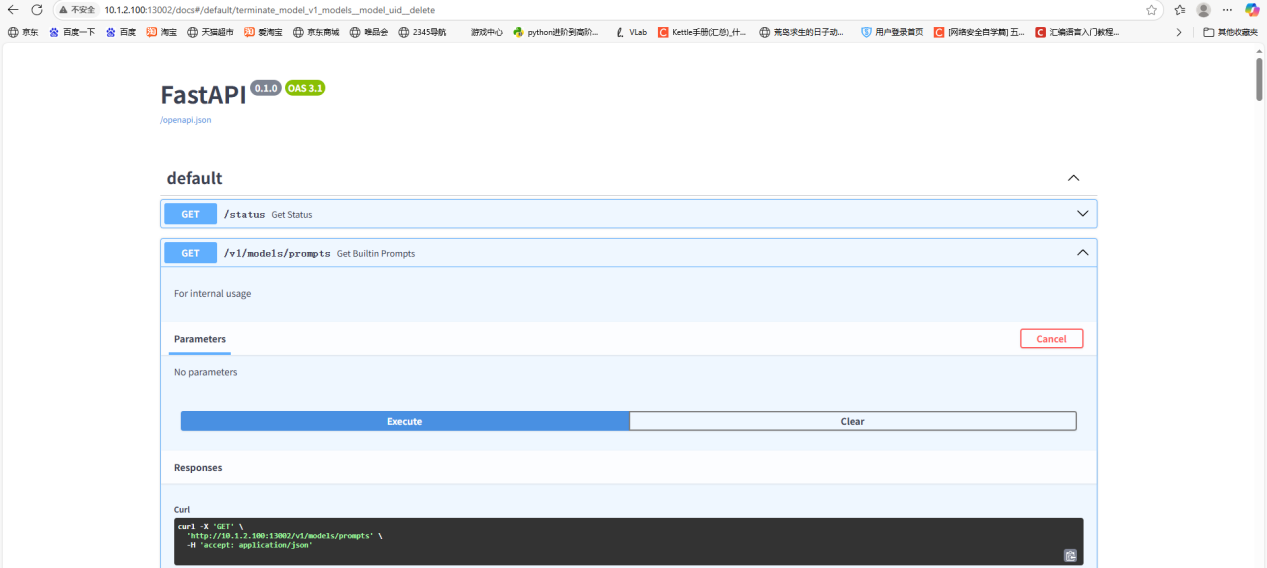
1. 默认配置情况下，本地运行Xinference的命令如下：

xinference-local --host 0.0.0.0 --port 9997

当前IP为10.1.2.100



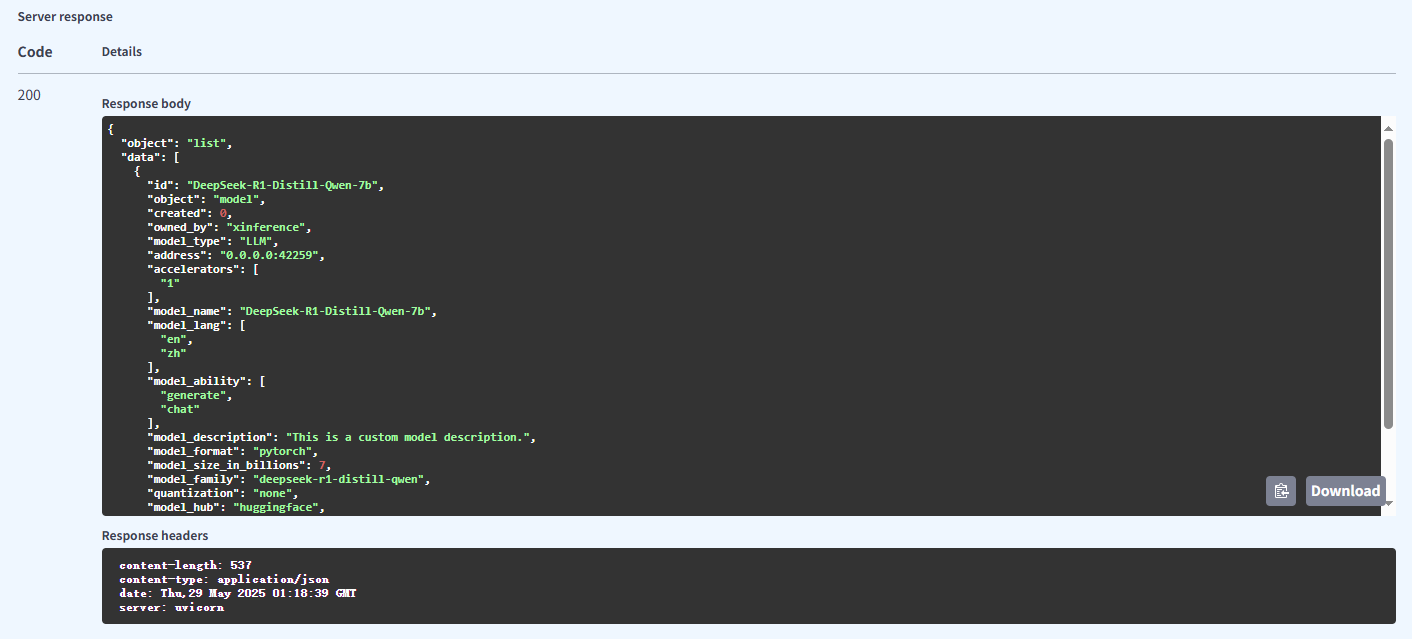
1. 访问http://10.1.2.100:13002/docs#，也可通过API接口调用工具实现下面步骤



1. 查看当前模型列表



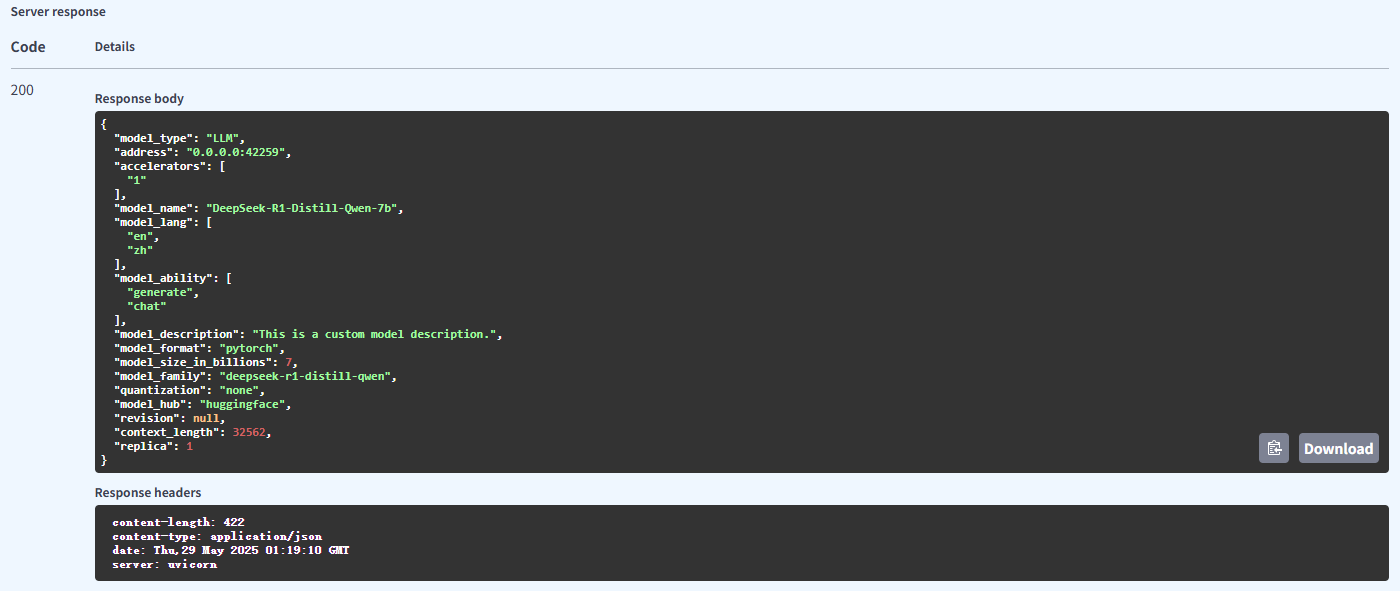
得到结果如下：



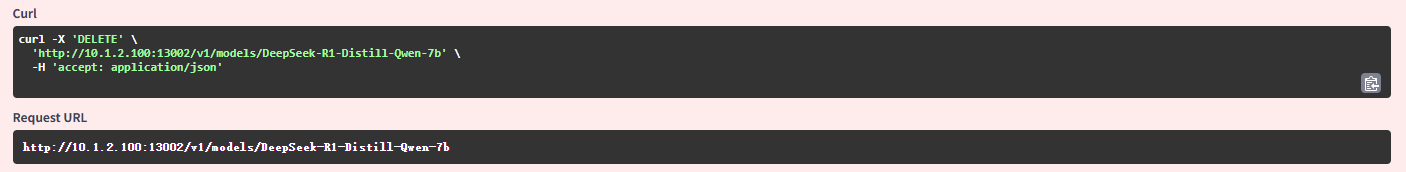
1. 访问模型相关描述



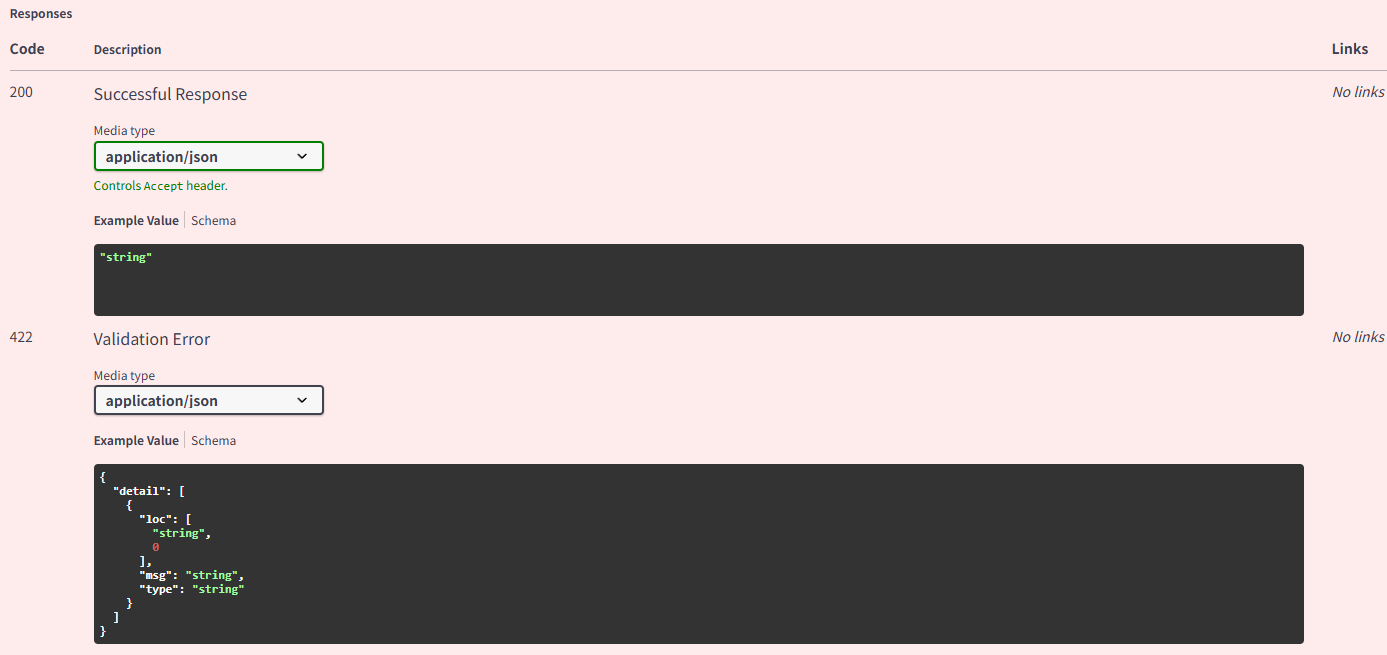
得到如下：



1. 删除指定模型



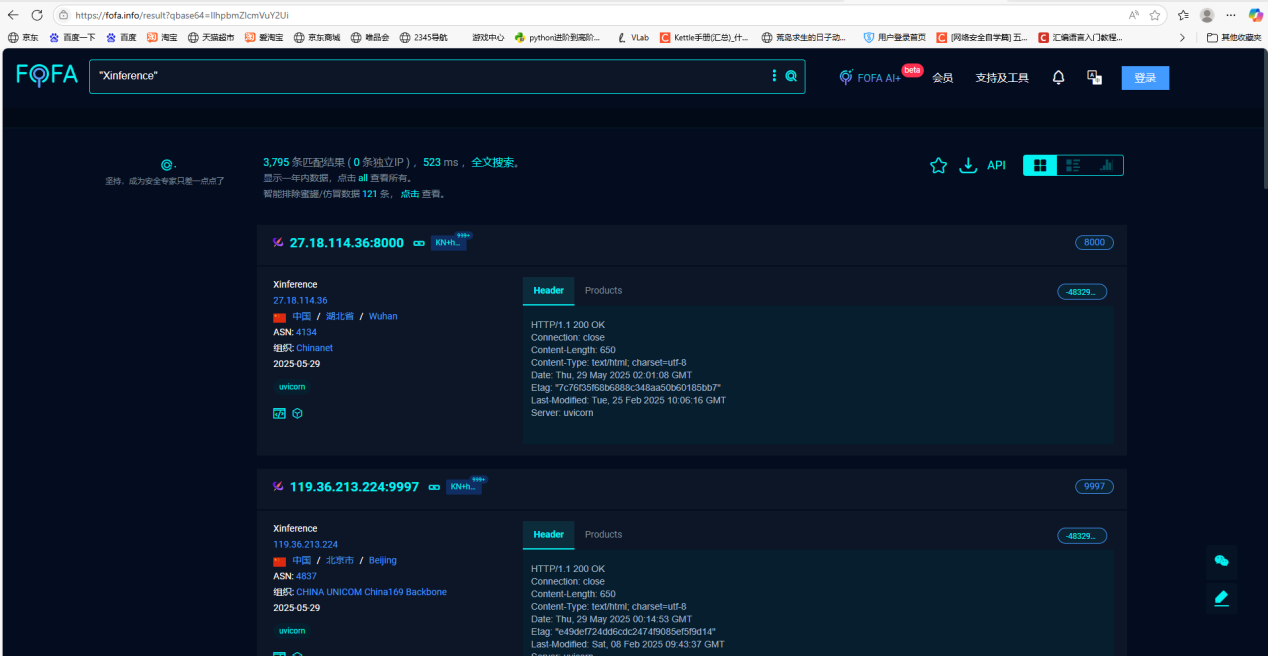
得到



进一步验证模型界面，确实该模型已经被删除

## 资产收集

通过"Xinference"命令，发现公网暴露资产不低于3795家，且大多数均可访问。



## 漏洞修复建议

1. 未必要避免将Xinference暴露在公网
2. 修改默认配置auth\_config\_file的值，避免为空；或设置其他权限控制方法。