



The ML(Ops) Security Landscape



the tone of this presentation and the memes used are not intended to offend anyone, especially the developers of the software to which the presentation relates

some of the misconfigurations/vulnerabilities discussed today are occurring “in the wild” - please keep in mind that using them to get access to the API keys/other people’s data is illegal

whoami

- My name is Mikołaj
- I work at **Ardoq** as a Software Security Analyst
- I blog on AI security at **Hackstery.com**
- I contribute to the **OWASP Top10 for Machine Learning**

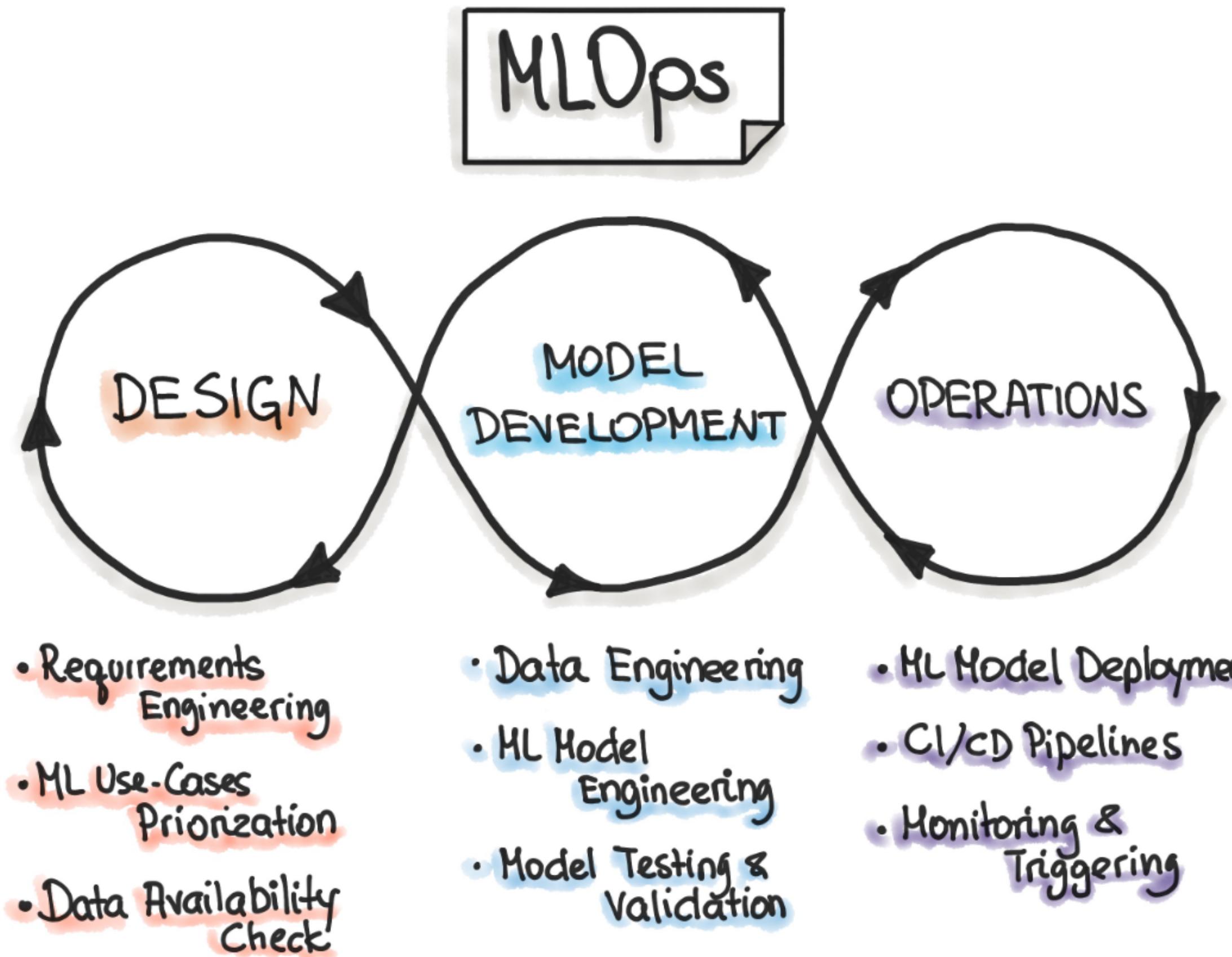


What will we talk about?

- What is MLOps/LLMOps?
- Vulnerabilities and misconfigurations in MLOps stack
- What are the OWASP initiatives for the security of AI/ML/LLMs?



MLOps



The complete MLOps process includes three broad phases of "*Designing the ML-powered application*", "*ML Experimentation and Development*", and "*ML Operations*".

source: <https://ml-ops.org/content/mlops-principles>

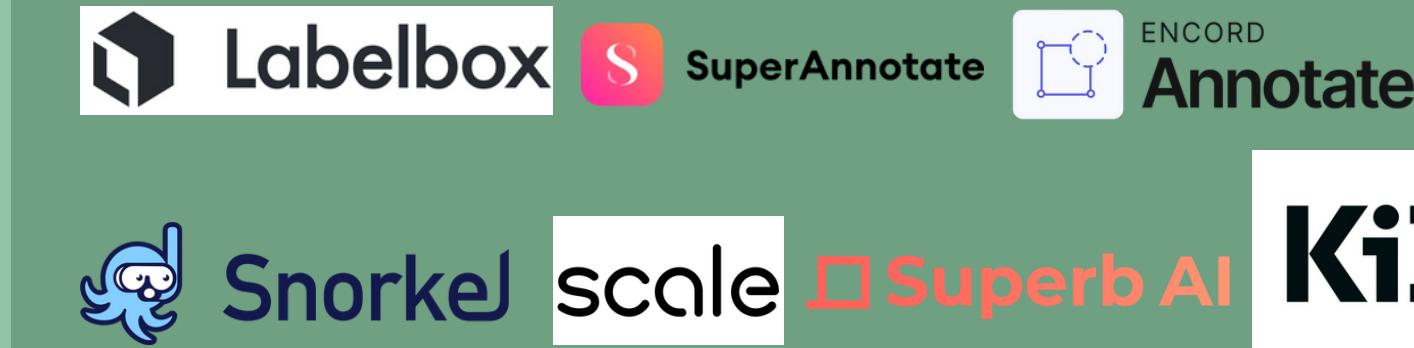


MLOps market map

MLOps platforms



Data labeling & annotation



Experiment tracking & model management



Model hubs



Workflow orchestration



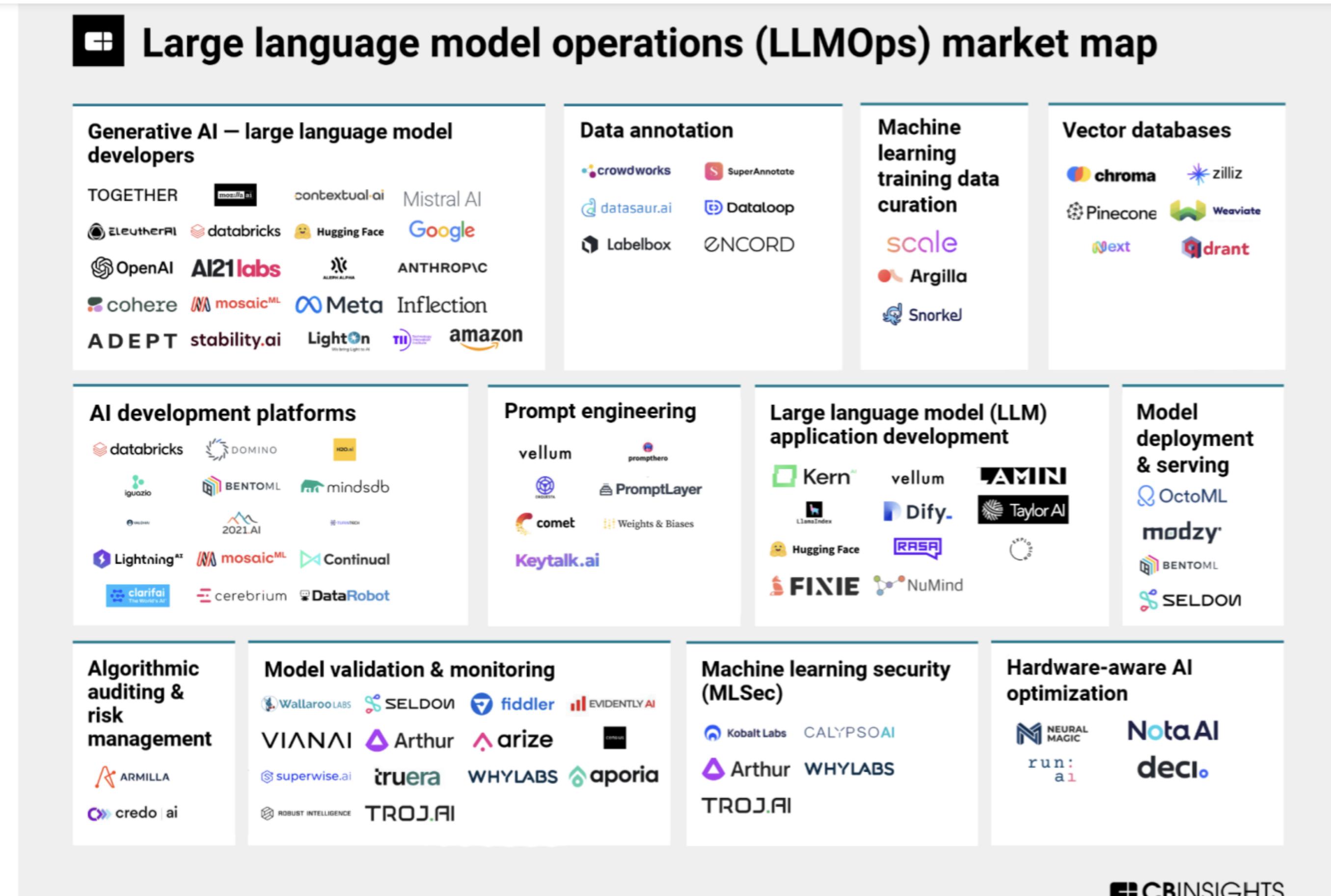
Model deployment



and many more...



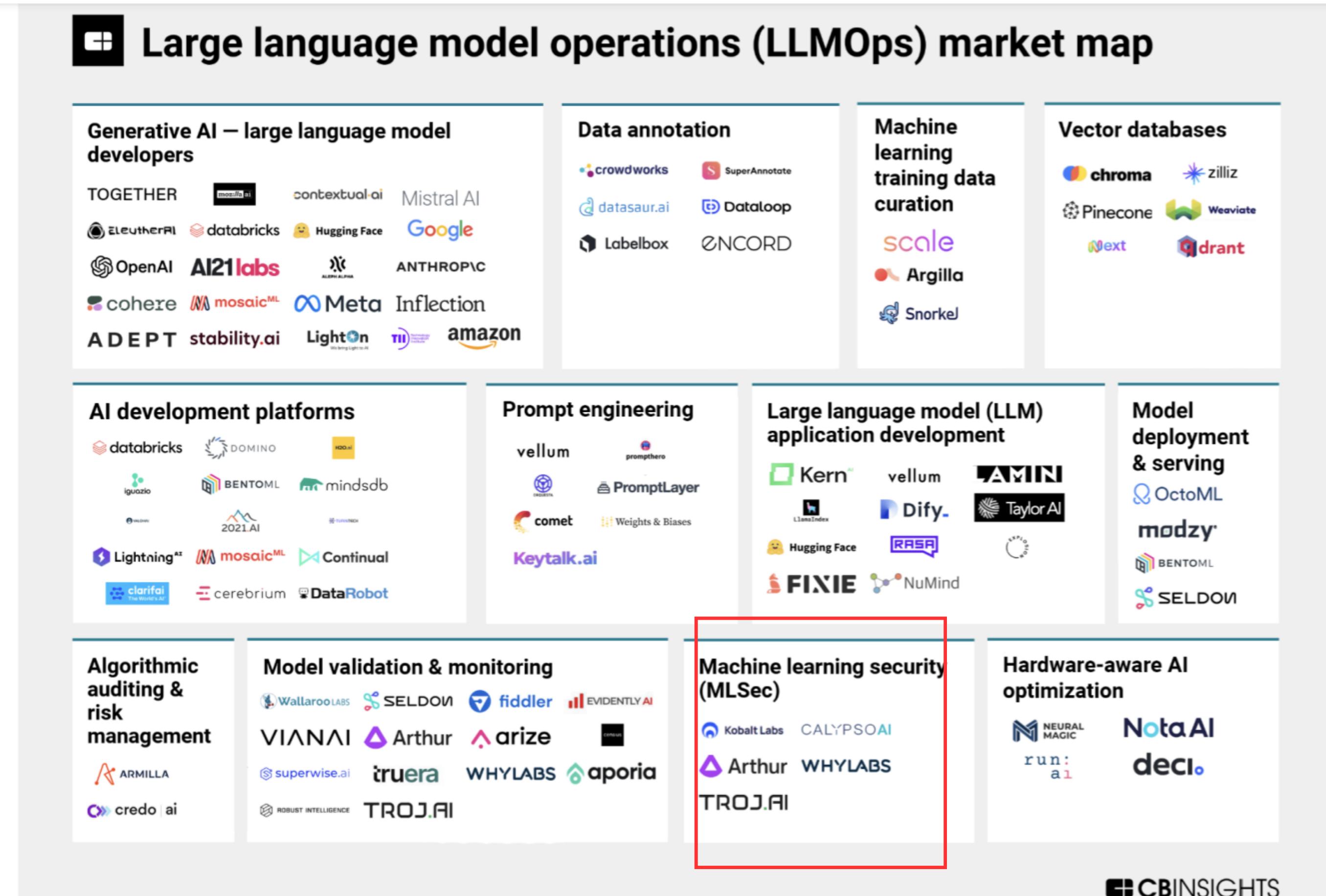
There's LLMOps market as well:



CB INSIGHTS

source: <https://www.cbinsights.com/research/large-language-model-operations-llops-market-map/>

There's LLMOps market as well:



source: <https://www.cbinsights.com/research/large-language-model-operations-llops-market-map/>

What's the problem with MLOps/LLMops?

- (sometimes) developers of these tools are not cybersecurity-aware
- tools are developed **too quickly**, resulting in lower security level
(you need to compromise something for the higher pace of development)



What's the problem with MLOps/LLMOps?

MLOps devs when you ask them about secret management



I've never met this man in my life.



title update:



The ML(Ops) Security Landscape





how to save a few dollars every month
by dumpster diving for OpenAI API keys



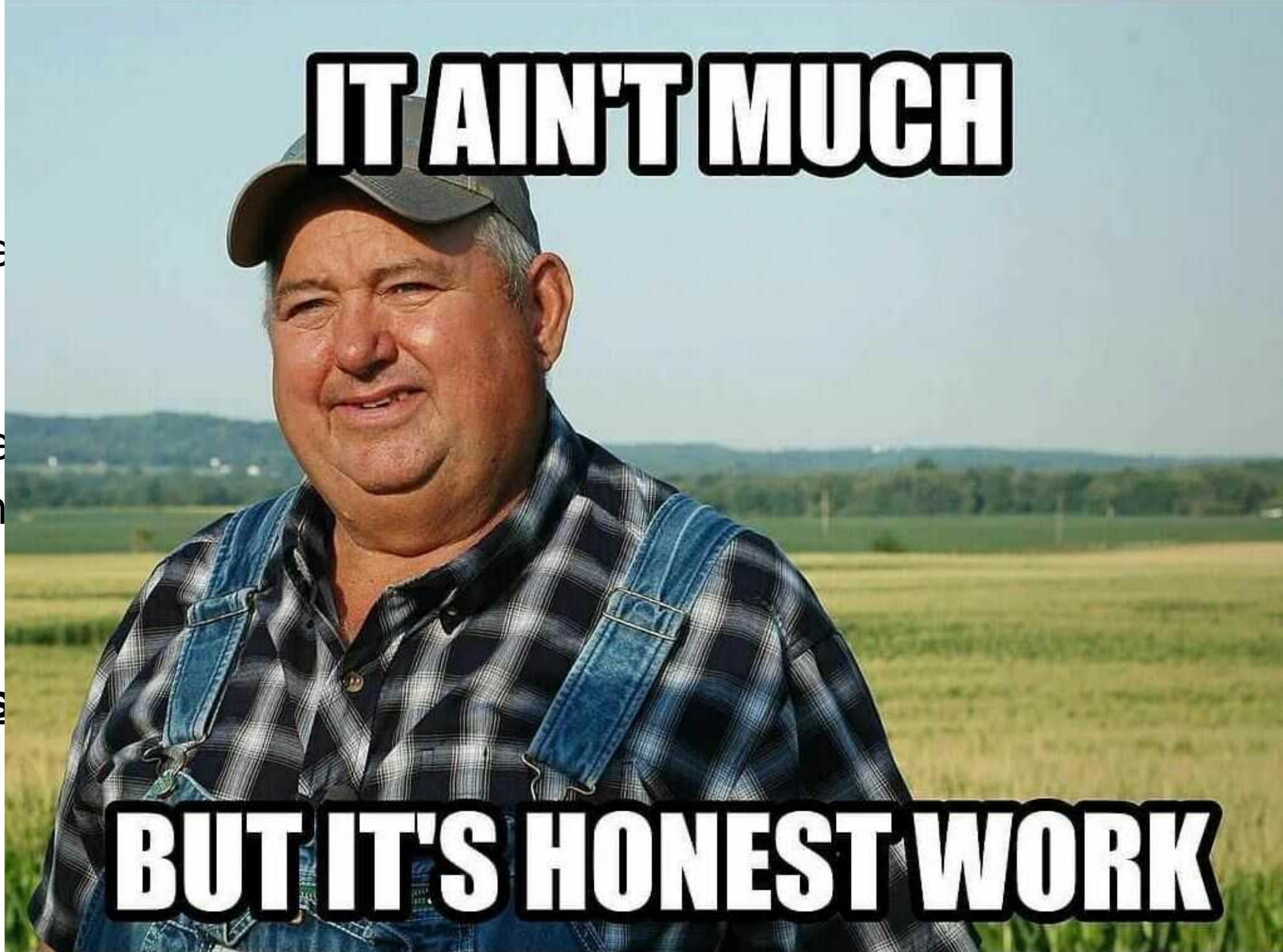
my approach

- I wanted to check **as many MLOps tools as possible** and verify, if some data is exposed “in the wild”
- I wanted to find the real “**low hanging fruits**”, so my target was the misconfigurations that you don’t have to actively exploit
- Due to the “quantity over quality” approach, I decided to use “**Awesome MLOps**” and “**Awesome LLMOps**” lists + **Shodan**



my approach

- I wanted to make sure the data is correct
- I wanted to avoid any misconceptions
- Due to the lack of MLOps



f some
ne
esome

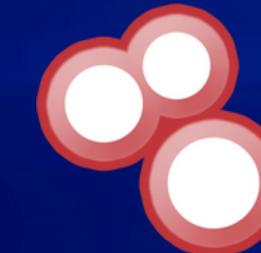


What is Shodan



Google

surface web



SHODAN



censys

deep web

torchl

dark web

What is Shodan

you can discover the most strange things connected to the Internet using Shodan

TOTAL RESULTS

25

TOP COUNTRIES



Hong Kong	11
Singapore	5
China	3
Netherlands	3
United States	3
More...	

TOP PORTS

443	12
80	10
3000	1
8005	1
8443	1
More...	

cloud

View Report

Download Results

Historical Trend

View on Map

Access Granted: Want to get more out of your existing Shodan account? Check

Smart Toilet Seat [🔗](#)

47.117.125.25

www.smarttoiletseat.net

smarttoiletseat.net

[Aliyun Computing Co., LTD](http://Aliyun%20Computing%20Co.,%20LTD)

China, Shanghai



SSL Certificate

Issued By:

| - Common Name:

R3

| - Organization:

[Let's Encrypt](#)

Issued To:

| - Common Name:

smarttoiletseat.net

Supported SSL Versions:

[TLSv1](#), [TLSv1.1](#), [TLSv1.2](#),

[TLSv1.3](#)

Diffie-Hellman Fingerprint:

[RFC3526/Oakley Group 14](#)

IPPE Buddy - smart toilet assistant - IPPE [🔗](#)

164.92.151.165

www.buddy.ippe.eu

home.ippe.eu

www.home.ippe.eu

ipee.eu

about.ippe.eu

[DigitalOcean, LLC](http://DigitalOcean.LLC)

Netherlands, Amsterdam



SSL Certificate

Issued By:

| - Common Name:

R3

| - Organization:

[Let's Encrypt](#)

Issued To:

| - Common Name:

ipee.eu

Supported SSL Versions:

[TLSv1.2](#), [TLSv1.3](#)



How can you use Shodan to find exposed MLOps software?

- using the Shodan dorks (*filters*), fortunately usually MLOps software has its name in the application's website title, so it is as easy as:

`http.html:"zenml"`

`http.html:"flowise"`

`http.html:"omniboard"`

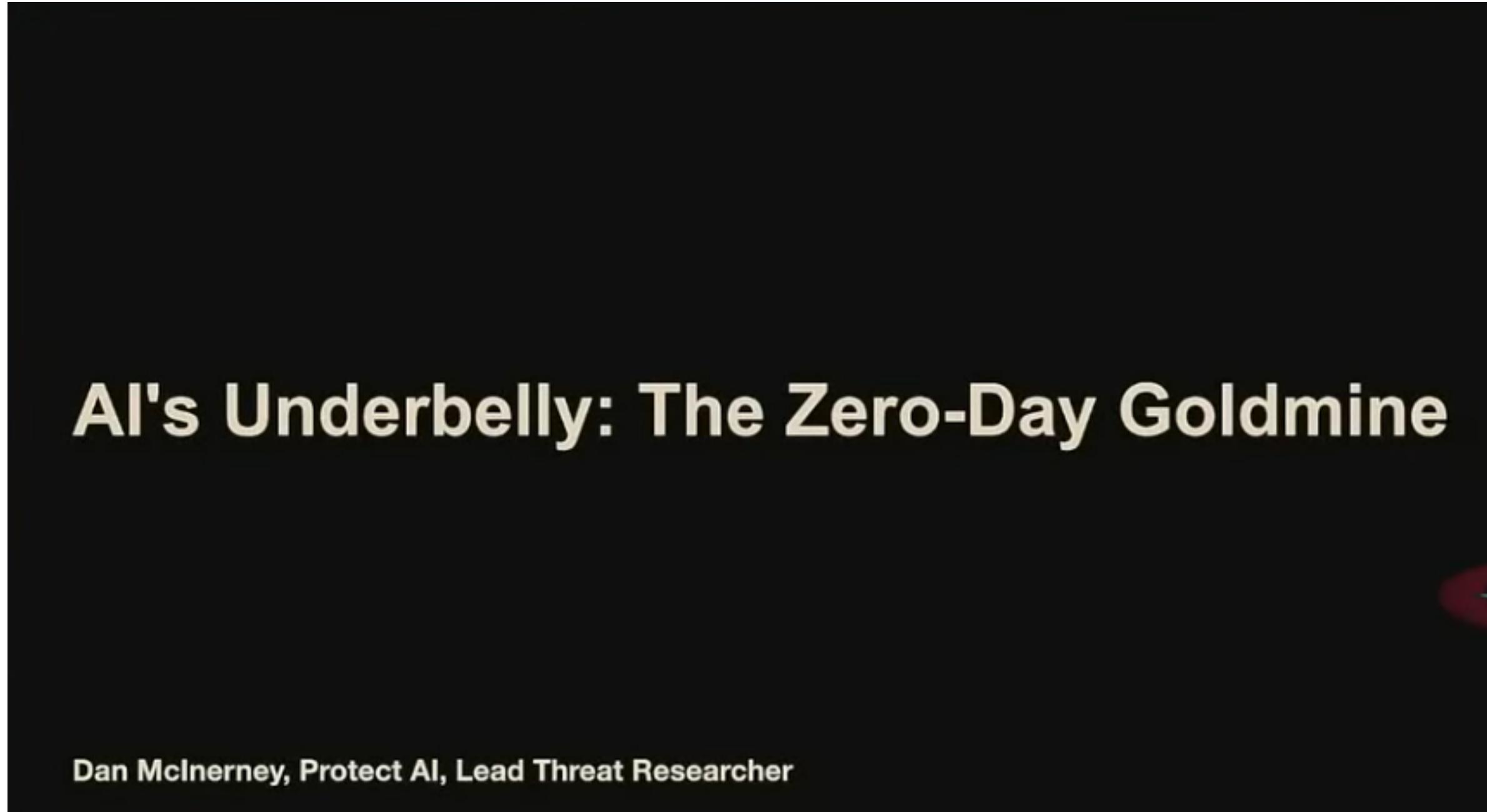
`http.html:"Prefect Server"`

▪ □ ▢ ▤ ▦ ▨ ▩ ▪ ▫ ▬ ▭ ▮



Kudos for inspiration

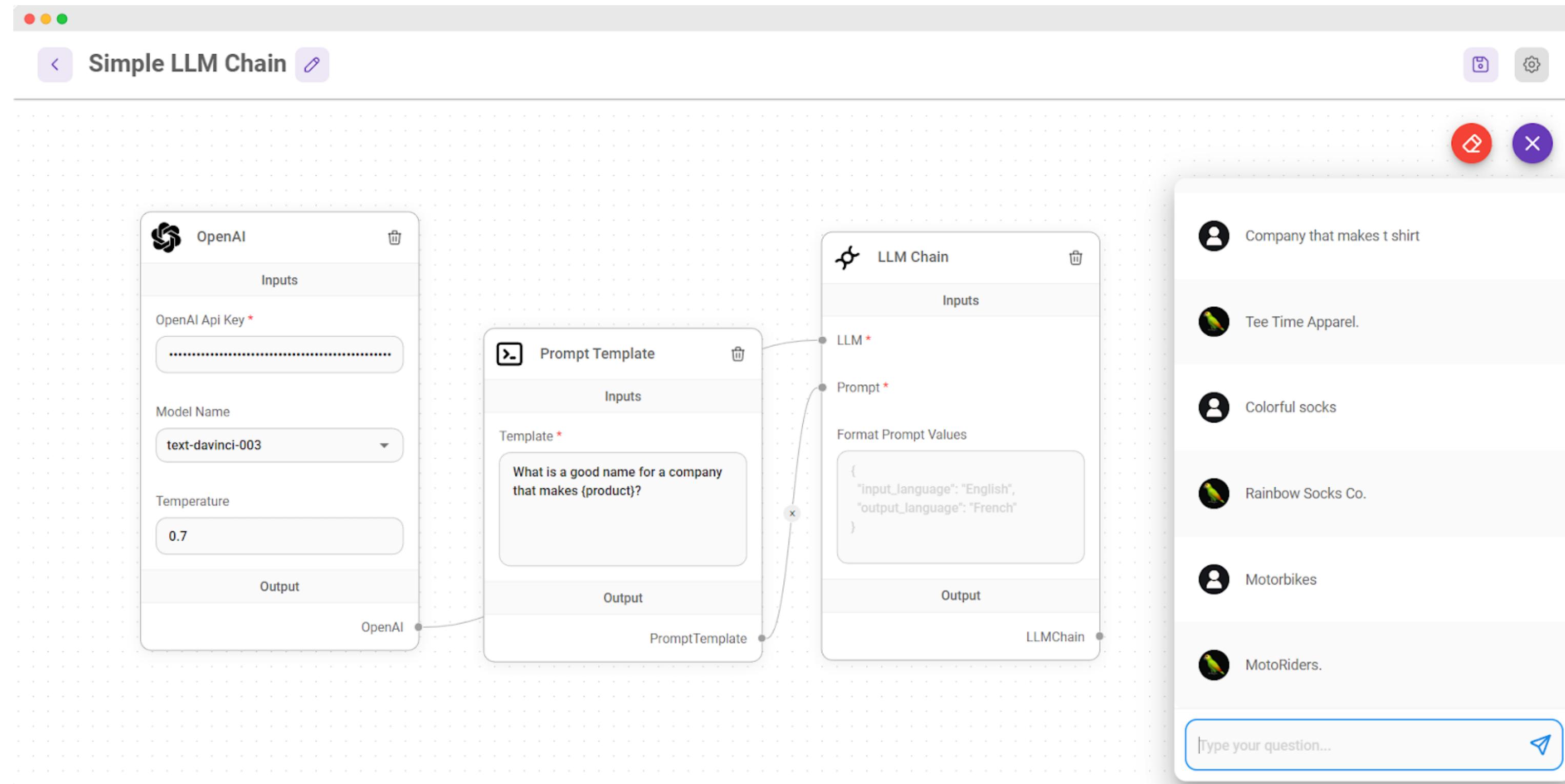
- Dan McInerney from ProtectAI @ ROOTCON



link to the talk: <https://www.youtube.com/watch?v=e3ybnXjtplc>



Case I: Flowise & Langflow



Flowise - anti-security features

- **lack of authentication** by default
- two ways of managing secrets
- in one of the cases, you could get plaintext keys from the server (**fixed**)
- **weird API endpoints**, i.e. endpoint for getting all of the conversation data and history



http.title:"flowise"

TOTAL RESULTS
1,121

TOP COUNTRIES



United States	455
Germany	136
Singapore	70
Netherlands	60
Australia	47
More...	

TOP PORTS	
3000	706
443	259
80	97
9000	5
5000	4
More...	

[View Report](#) [Download Results](#) [Historical Trend](#) [View on Map](#)

Access Granted: Want to get more out of your existing Shodan account? Check out [everything you have access to](#).

Flowise - Low-code LLM apps builder ↗

94.241.174.232
n.i-setka.ru
TimeWeb Ltd.
Netherlands, Amsterdam

eol-product

SSL Certificate

Issued By:
|- Common Name:
R3
Issued To:
|- Organization:
Let's Encrypt

Supported SSL Versions:
TLSv1.2, TLSv1.3

HTTP/1.1 200 OK
Server: nginx/1.18.0 (Ubuntu)
Date: Sun, 26 Nov 2023 23:40:50 GMT
Content-Type: text/html; charset=UTF-8
Content-Length: 2116
Connection: keep-alive
X-Powered-By: Express
Access-Control-Allow-Origin: *
Accept-Ranges: bytes
Cache-Control: public, max-age=0
Last-Modified: ...

Flowise - Low-code LLM apps builder ↗

104.131.180.158
DigitalOcean, LLC
United States, Clifton

cloud

HTTP/1.1 200 OK
X-Powered-By: Express
Access-Control-Allow-Origin: *
Accept-Ranges: bytes
Cache-Control: public, max-age=0
Last-Modified: Wed, 09 Aug 2023 23:08:22 GMT
ETag: W/"844-189dc8eee70"
Content-Type: text/html; charset=UTF-8
Content-Length: 2116
Date: Sun, 26 Nov 2023 23:27:51 GM...

Flowise - Low-code LLM apps builder ↗

38.60.204.248

HTTP/1.1 200 OK



or http.title:"flowise" port:3000



/api/v1/database/export

Pretty Raw Hex

```
1 GET /api/v1/database/export HTTP/1.1
2 Host: [REDACTED]
3
4
5
6
7 Sec-Ch-Ua-Platform: "Linux"
8 Sec-Fetch-Site: same-origin
9 Sec-Fetch-Mode: cors
10 Sec-Fetch-Dest: empty
11 Referer: [REDACTED]
12 Accept-Encoding: gzip, deflate, br
13 Accept-Language: en-US,en;q=0.9
14 Connection: close
15
16
```

Pretty Raw Hex Render

```
{  
  "id": "bf[REDACTED]",  
  "role": "userMessage",  
  "chatflowid": "[REDACTED]",  
  "content": "Bonjour",  
  "sourceDocuments": null,  
  "createdDate": "[REDACTED"]},  
,  
{  
  "id": "76[REDACTED]",  
  "role": "apiMessage",  
  "chatflowid": "e[REDACTED]",  
  "content": "

Bien  
us aujourd'hui",  
  "sourceDocuments": null,  
  "createdDate": "[REDACTED"]},  
,  
{  
  "id": "[REDACTED]",  
  "role": "userMessage",  
  "chatflowid": "e[REDACTED]",  
  "content": "déjà, [REDACTED] ? ",  
  "sourceDocuments": null,  
  "createdDate": "2023-03-07T09:49:47.000Z"},  
,  
{  
  "id": "7[REDACTED] 2023-03-07T09:49:47.000Z",  
  "role": "apiMessage",  
  "chatflowid": "[REDACTED]",  
  "content": "

Je suis [REDACTED]  
longem[REDACTED]

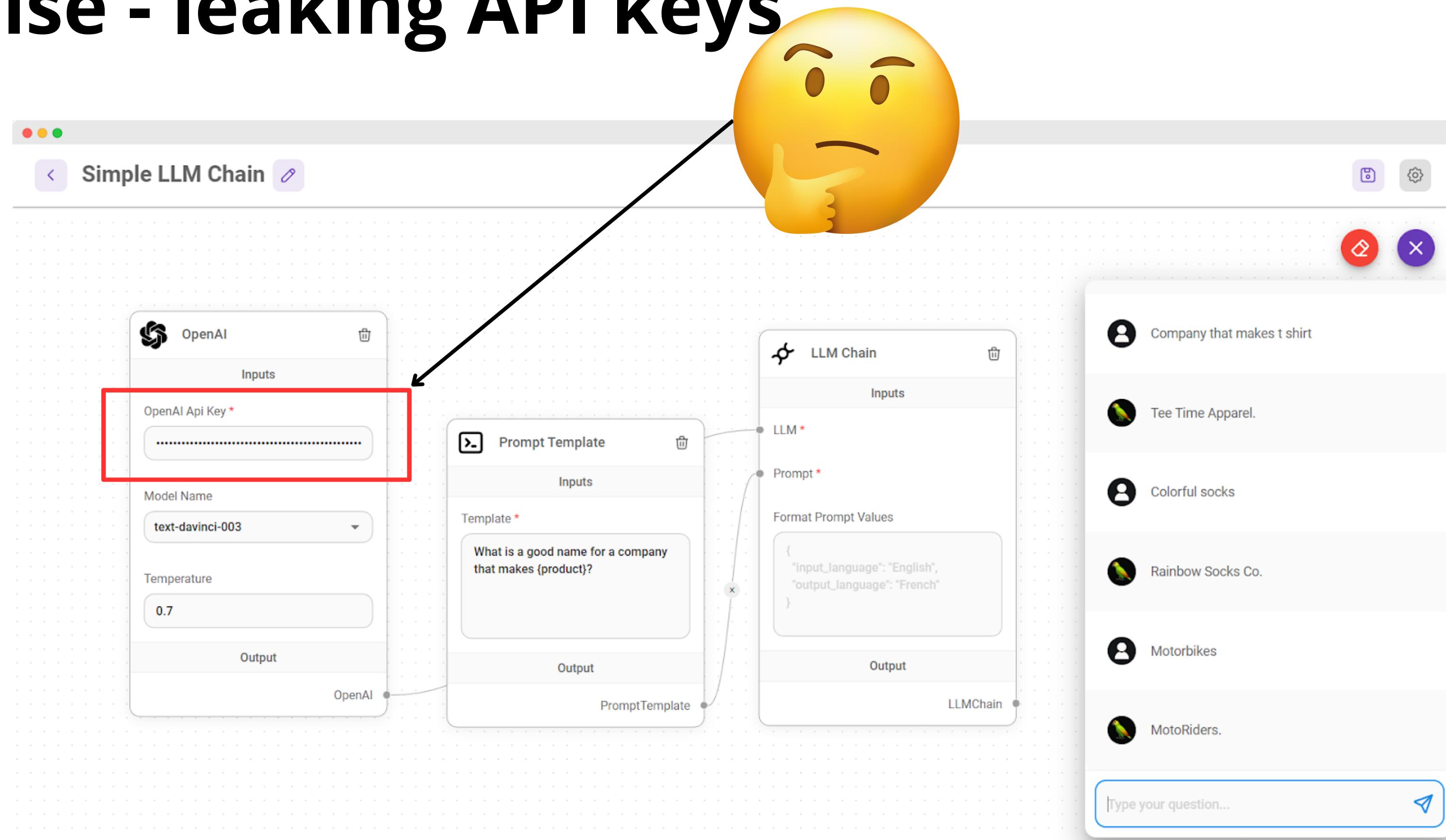


[REDACTED] à l'heure. Comment puis-je vous aider

",  
  "sourceDocuments": null,  
  "createdDate": "[REDACTED"]}

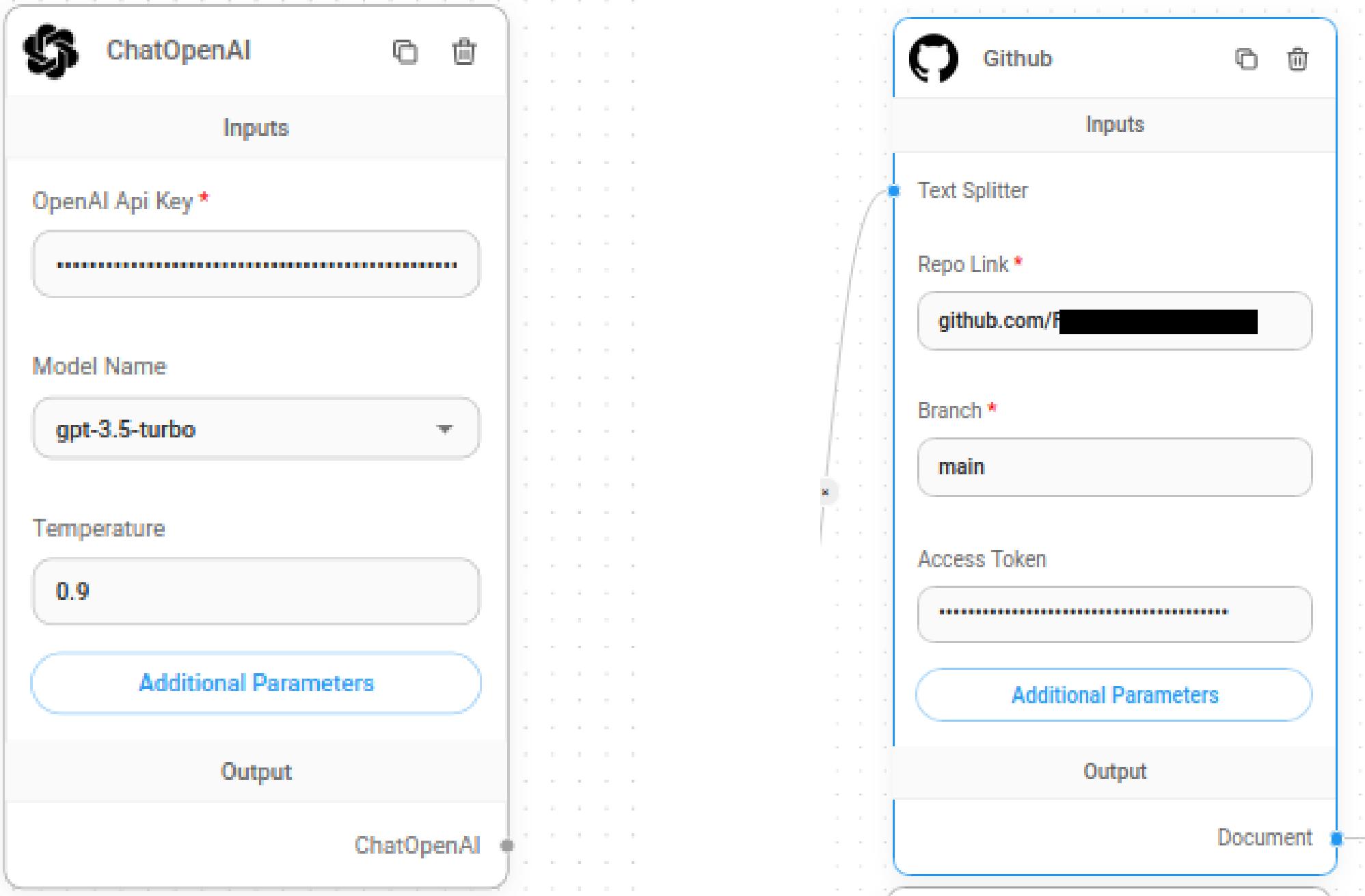

```

Flowise - leaking API keys



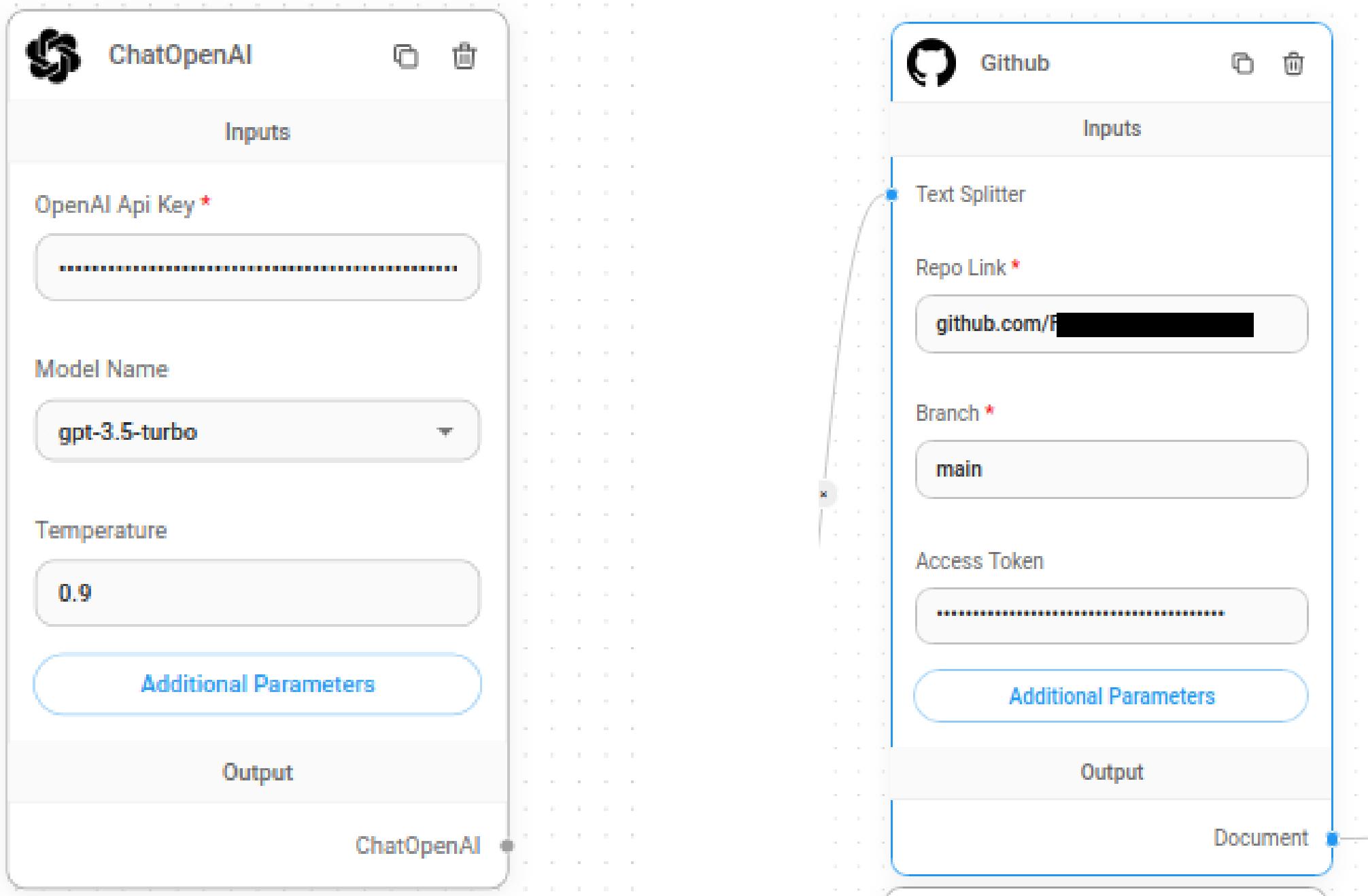
This example comes from the Flowise docs

Flowwise - leaking API keys



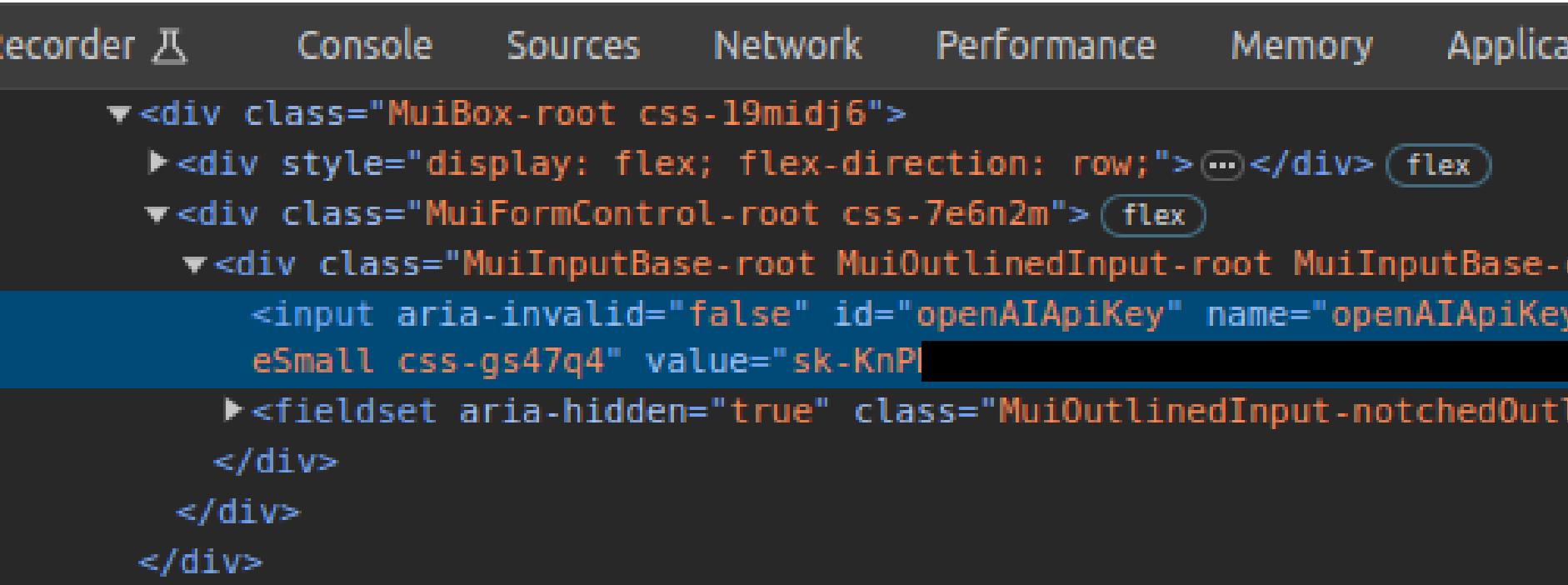
Flowwise - leaking API keys

If we just inspect an element in the browser it would be too easy, right?



Flowise - leaking API keys

If we just inspect an element in the browser it would be too easy, right? **No**

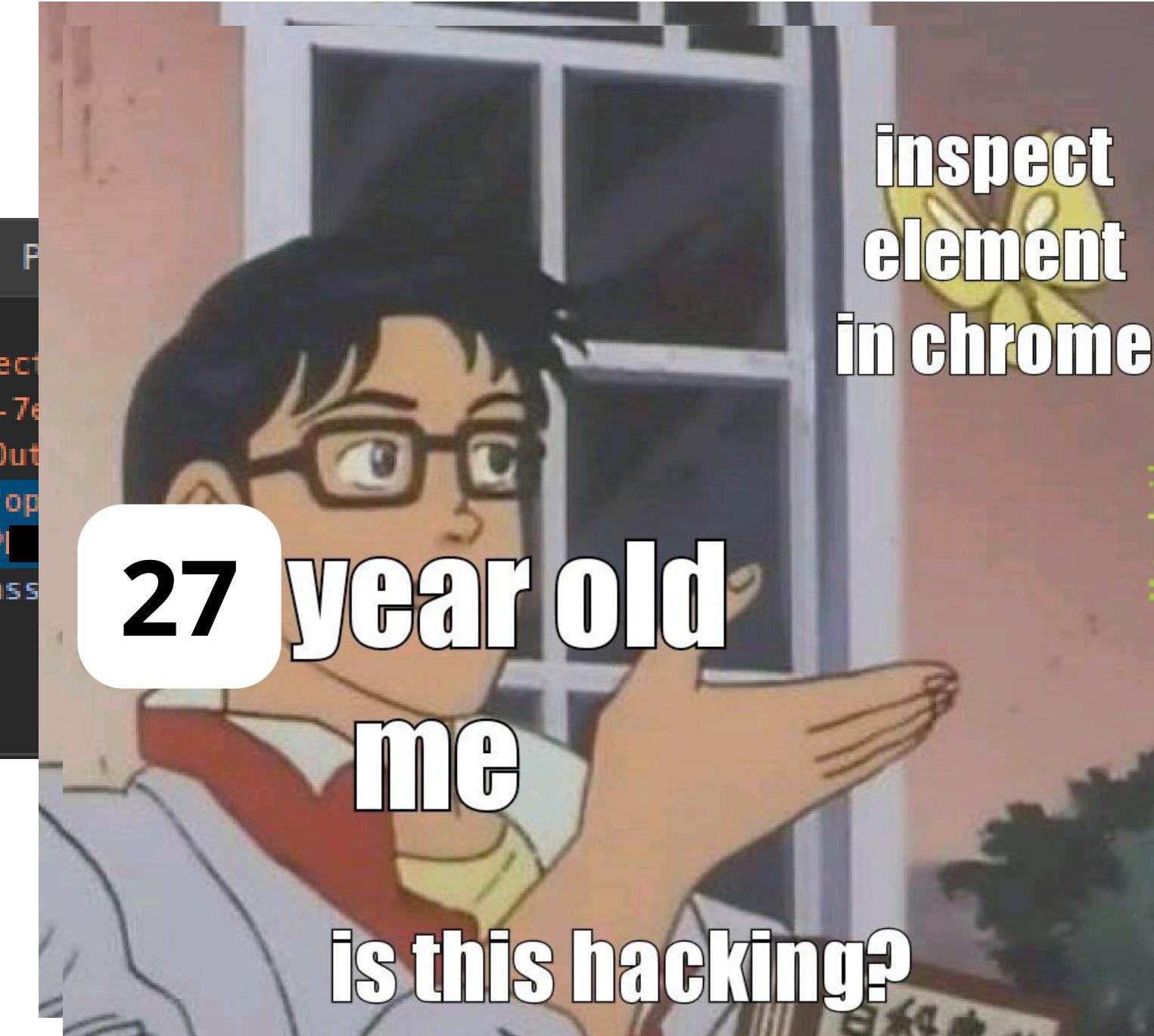


The screenshot shows the Network tab of a browser's developer tools. A single request is listed:

```
POST https://api.openai.com/v1/engines/  
Content-Type: application/json  
Content-Length: 1033  
  
{  
  "name": "My Engine",  
  "type": "text",  
  "description": "A large language model trained by OpenAI.",  
  "parameters": {  
    "max_tokens": 100,  
    "temperature": 0.5,  
    "top_p": 1.0,  
    "frequency_penalty": 0.0,  
    "presence_penalty": 0.0,  
    "stop": null  
  },  
  "model": "text-davinci-003",  
  "n": 1,  
  "best_of": 1,  
  "logprobs": 0,  
  "echo": false,  
  "allow_retries": true,  
  "max_retry": 3,  
  "timeout": 10000,  
  "presence_penalty_weight": 0.0,  
  "frequency_penalty_weight": 0.0,  
  "seed": null  
}
```

Flowise - leaking API keys

```
Recorder ▾      Console      Sources      Network      P  
      ▼<div class="MuiBox-root css-19midj6">  
        ▶<div style="display: flex; flex-direction:  
      ▼<div class="MuiFormControl-root css-7e  
        ▶<div class="MuiInputBase-root MuiOutl  
          <input aria-invalid="false" id="op  
            eSmall css-gs47q4" value="sk-KnPi  
        ▶<fieldset aria-hidden="true" class="  
          </div>  
        </div>  
      </div>
```



is this hacking?

I informed some companies and individuals, some ghosted me, some said “thank you”

Hey Mikolaj,

Thanks for the highlight! Even though the keys are there I don't think people can copy them as Flowise doesn't allow it.

Curious if that's possible and how you found out about this.

Rhar I like your newsletter "Real Threats of Artificial Intelligence" keep up the good work!

flowise not allowing it:



Hey Mi

Thanks

Curious

Rhar I li

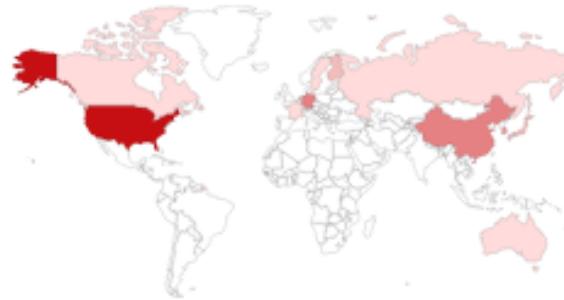
hem as Flowise doesn't allow it.

Langflow

TOTAL RESULTS

63

TOP COUNTRIES



United States 30

China 6

Germany 6

Korea, Republic of 4

Singapore 4

[More...](#)

TOP PORTS

3000 28

80 13

443 9

8080 7

6009 1

[More...](#)

TOP ORGANIZATIONS

Amazon Technologies Inc. 14

View Report

Download Results

Historical Trend

View on Map

Access Granted: Want to get more out of your existing Shodan account? Check out [everything you can do with Shodan](#).

Langflow [🔗](#)

20.121.52.131
Microsoft Corporation
 United States, Ashburn



cloud

HTTP/1.1 200 OK
date: Mon, 27 Nov 2023 04:12:15 GMT
server: unicorn
content-type: text/html; charset=utf-8
content-length: 724
last-modified: Tue, 25 Jul 2023 00:13:54 GMT
etag: 46ff59872e5a8cd55db1e1cf25dbaf5

LangFlow [🔗](#)

39.107.240.252
Aliyun Computing Co., LTD
 China, Beijing

eol-product

HTTP/1.1 200 OK
Server: nginx/1.20.1
Date: Mon, 27 Nov 2023 00:51:16 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 622
Connection: keep-alive
last-modified: Thu, 10 Aug 2023 02:25:41 GMT
etag: 3736508e8c41a67d18593e88e768a85f

Langflow [🔗](#)

185.130.226.250
HOSTKEY B.V.
 Netherlands, Amsterdam

HTTP/1.1 200 OK
Access-Control-Allow-Origin: *
Content-Type: text/html
Cache-Control: no-cache
Etag: W/"379-ILZRPDD732MID05zG2029R/RyU8"
Date: Sun, 26 Nov 2023 21:35:21 GMT
Connection: keep-alive
Keep-Alive: timeout=5
Content-Length: 889



Langflow

Langflow is a frontend for Langchain (similar to Flowise)

The screenshot shows the Langflow interface, which is a visual workflow builder for Langchain. The left sidebar lists various components: Agents, Chains, Loaders, Embeddings, LLMs, Memories, Output Parsers, Prompts, Retrievers, Text Splitters, Toolkits, Tools, Utilities, Vector Stores, and Wrappers. The main area displays a workflow graph with three main nodes:

- ChatOpenAI**: A node for wrapping OpenAI Chat large language models. It has fields for Max Tokens (Type an integer number), Model Name (gpt-3.5-turbo-0613), OpenAI API Base (Type something...), and OpenAI API Key (redacted). It also includes Temperature (0.7) and a toolbar with icons for delete, copy, file, and key.
- Search**: A search engine node. It has a Serpapi Api Key field and a Search button.
- ZeroShotAgent**: An agent construction node. It requires an LLM (selected as ChatOpenAI), Tools (selected as Search), and Prefix (Answer the following questions as best you can). It also has a Suffix field.

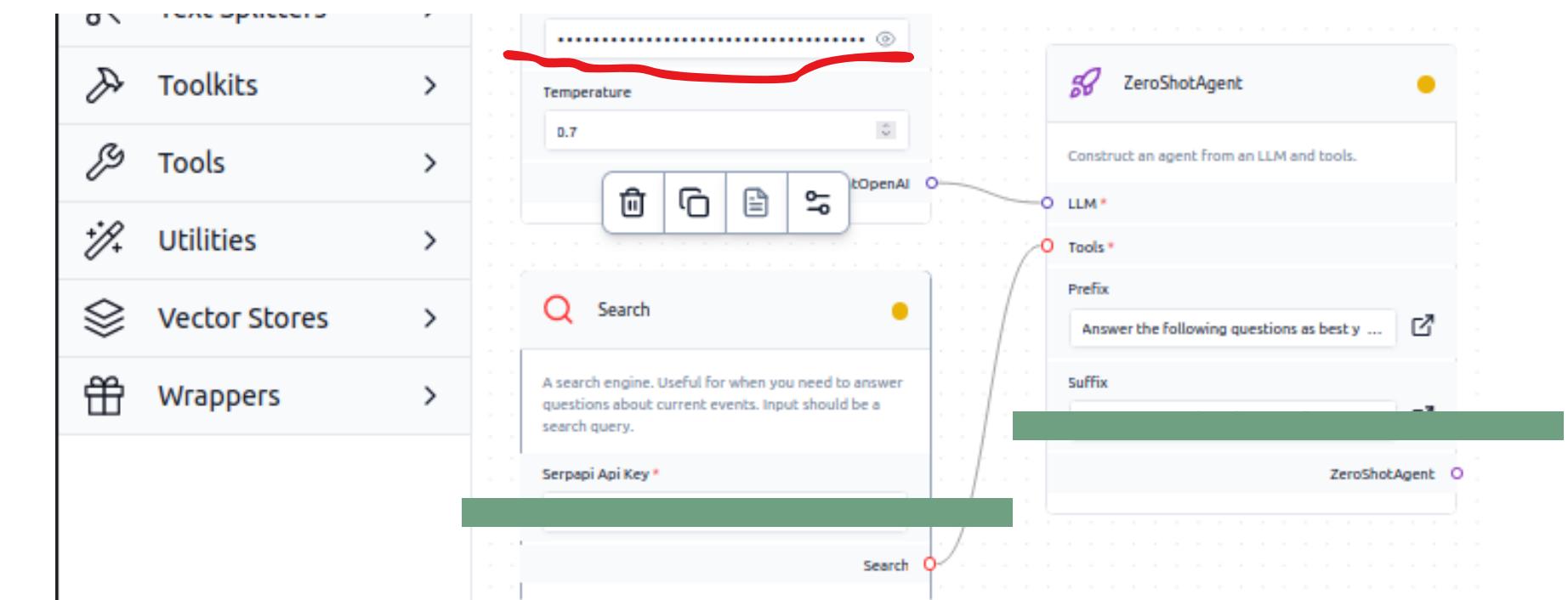
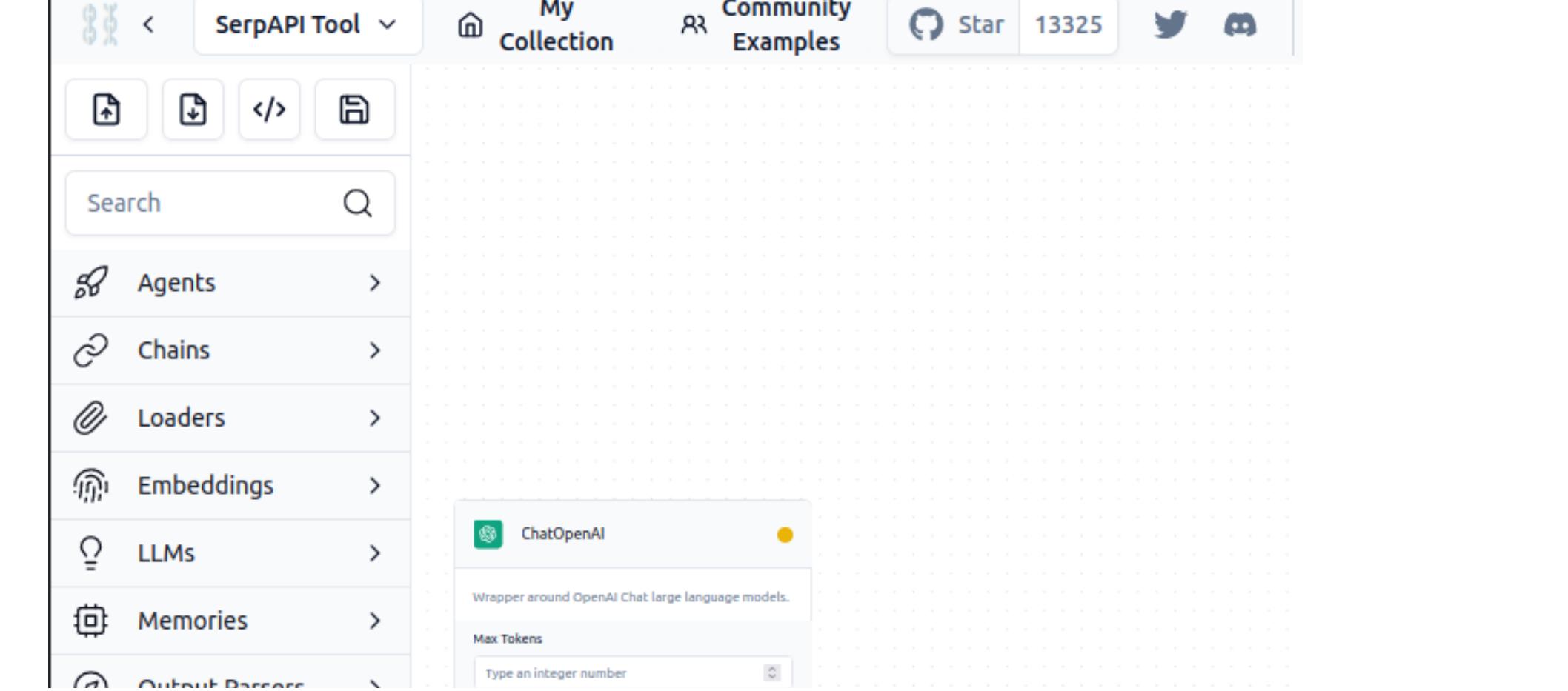
Connections between the nodes are shown as lines: an arrow from ChatOpenAI to Search, and another from Search to ZeroShotAgent. The ZeroShotAgent node is highlighted with a green bar at the bottom.



Langflow

Reply from the developers:

The way it works at the moment, the API key does not come back from the backend. It is sent from the browser to the backend and used there.



Langflow

...and we're back to the problem of lack of security awareness:

```
GET /api/v1/flows/ HTTP/1.1
Host: xxxxxx:3000
Accept: application/json, text/plain, */*
Authorization: Bearer eyxxxxxxxxxxxxxU
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Referer: http://xxxxxx:3000/
Accept-Encoding: gzip, deflate, br
Accept-Language: en-US,en;q=0.9
Cookie: access_tkn_lflw=eyJxxxxxx; refresh_tkn_lflw=auto
Connection: close
```

```
HTTP/1.1 200 OK
Access-Control-Allow-Origin: *
date: Thu, 16 Nov 2023 17:52:37 GMT
server: uvicorn
content-length: 57204
content-type: application/json
connection: close

[...]
"openai_api_key": {"required": false, "placeholder": "", "show": true, "multiline": false,
                    "value": "sk-xxxxxxxxxxxxxx0R",
                    "password": true, "name": "openai_api_key", "display_name": "OpenAI API Key", "advanced": false, "dynam
[...]
```

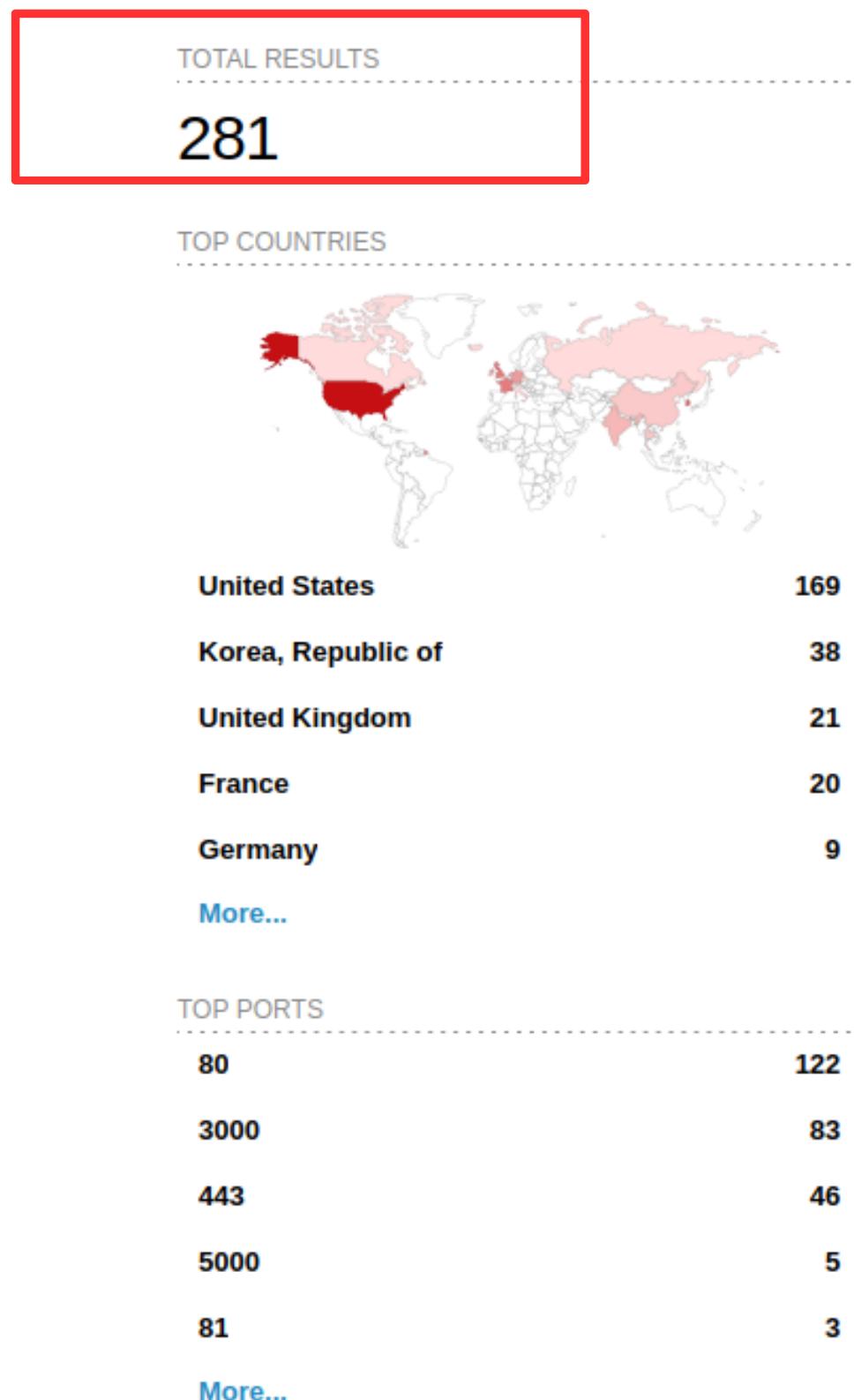


case II

- let's say you're too lazy to use OpenAI API and you want to access the model directly through the exposed MLOps software



case II: BentoML



[View Report](#) [Download Results](#) [Historical Trend](#) [View on Map](#)

Access Granted: Want to get more out of your existing Shodan account? [Check it out!](#)

BentoML Prediction Service 🔗	
35.172.210.62	HTTP/1.1 200 OK
ec2-35-172-210-62.compute-1.amazonaws.com	Date: Mon, 27 Nov 2023 00:14:01 GMT
Amazon Technologies Inc.	Content-Type: text/html; charset=utf-8
🇺🇸 United States, Ashburn	Content-Length: 2859
	Connection: keep-alive
	server: uvicorn
cloud	
BentoML Prediction Service 🔗	
18.234.8.154	HTTP/1.1 200 OK
ec2-18-234-8-154.compute-1.amazonaws.com	Date: Sun, 26 Nov 2023 23:33:49 GMT
Amazon Technologies Inc.	Content-Type: text/html; charset=utf-8
🇺🇸 United States, Ashburn	Content-Length: 2859
	Connection: keep-alive
cloud	server: uvicorn
BentoML Prediction Service 🔗	
3.123.234.211	HTTP/1.1 200 OK
ec2-3-123-234-211.eu-central-1.compute.amazonaws.com	date: Sun, 26 Nov 2023 23:26:03 GMT
A100 ROW GmbH	server: uvicorn
🇩🇪 Germany, Frankfurt am Main	content-length: 2945
	content-type: text/html; charset=utf-8
cloud	



BentoML - anti-security features

- **lack of authentication** by default (and in 90% of cases it's deployed without authentication)
- well documented **API access to the models** - both LLMs and other types of Machine Learning models
- supports **Swagger** and by default exposes SwaggerUI



case II: BentoML

Request body required

application/json

Examples:

[Modified value] ▾

```
{  
  "prompt": "In order to secure your MLops tools, you need to do the following:",  
  "model": "llama2-7b-chat-uncensored",  
  "max_tokens": 256,  
  "temperature": 0.7,  
  "logprobs": 1,  
  "top_p": 0.43,  
  "n": 1,  
  "stream": false  
}
```

Execute Clear



case II: BentoML

Server response

Code Details

200

Response body

```
{  
  "choices": [  
    {  
      "index": 0,  
      "text": "\n\n1. Implement a strong authentication mechanism to ensure that only authorized users can access the MLOps tools.\n2. Enforce strict access control policies to limit the access of users to only the necessary resources.\n3. Use encryption to protect sensitive data stored in the MLOps tools.\n4. Regularly update the MLOps tools with the latest security patches and vulnerability fixes.\n5. Conduct regular security audits to identify and address any potential security vulnerabilities.\n6. Implement a robust incident response plan to quickly respond to any security incidents.\n7. Train your employees on the importance of security and how to handle sensitive data securely.\n8. Use a secure network infrastructure to prevent unauthorized access to the MLOps tools.\n9. Use a secure hosting environment to ensure that the MLOps tools are protected from external threats.\n10. Implement a robust data governance framework to ensure that the data stored in the MLOps tools is managed securely.\nBy implementing these security measures, you can secure your MLOps tools and protect your organization from potential security threats.",  
      "logprobs": {  
        "text_offset": [  
          0,  
          ...  
        ]  
      }  
    }  
  ]  
}
```



case II: BentoML

Server response

Code Details

200

Response body

```
{  
    "choices": [  
        {  
            "index": 0,  
            "text": "\n\n1. Implement a strong authentication mechanism to ensure that only authorized users can access the MLOps tools.\n2. Enforce strict access control policies to limit the access of unauthorized users.\n3. Use encryption to protect sensitive data stored in the MLOps tools.\n4. Regularly update the MLOps tools to apply the latest security vulnerability fixes.\n5. Conduct regular security audits to identify and address any potential security vulnerabilities.\n6. Create an incident response plan to quickly respond to any security incidents.\n7. Train your employees on the importance of security and how to prevent data breaches.\n8. Use a secure network infrastructure to prevent unauthorized access to the MLOps tools.\n9. Implement a robust data governance system to ensure that sensitive data stored in the MLOps tools is managed securely.\n10. By implementing these security measures, you can secure your MLOps tools from potential security threats.",  
            "text_offset": [0]
```



case III

- but what if you wanted to get the output that would be easier to read? And maybe you wanted to compare multiple LLMs in one place? **Misconfigured MLOps tools got you covered!**



case III: Streamlit

TOTAL RESULTS
8,882

TOP COUNTRIES



United States	4,680
Germany	630
Ireland	623
China	335
India	319
More...	

TOP PORTS

80	4,360
443	2,980
8080	373
8443	108
8000	97
More...	

[View Report](#) [Download Results](#) [Historical Trend](#) [View on Map](#)

Access Granted: Want to get more out of your existing Shodan account? Check out our [pricing](#).

 **Streamlit** ↗

34.78.150.55
55.150.78.34.bc.googleusercontent.com
Google LLC
Belgium, Brussels
 cloud

HTTP/1.1 200 OK
Server: TornadoServer/6.1
Content-Type: text/html
Date: Mon, 27 Nov 2023 01:43:59 GMT
Accept-Ranges: bytes
Etag: "a6a3f16b289fba980673f9a25817dc6d1e1681a198b4"
Last-Modified: Tue, 05 Oct 2021...

 **Streamlit** ↗

45.113.234.202
vm-45-113-234-202.rc.cloud.unimelb.edu.au
The University of Melbourne
 Australia, Melbourne

HTTP/1.1 200 OK
Server: TornadoServer/6.3.2
Content-Type: text/html
Date: Mon, 27 Nov 2023 01:43:35 GMT
Accept-Ranges: bytes
Etag: "clea7740a2f00ab5ba65a775e46bc6720678138eb843"
Last-Modified: Wed, 12 Jul 20...

 **Streamlit** ↗

3.39.191.164
ec2-3-39-191-164.ap-northeast-2.compute.amazonaws.com
AWS Asia Pacific (Seoul) Region
 Korea, Republic of, Seoul


HTTP/1.1 200 OK
Server: TornadoServer/6.3.2
Content-Type: text/html
Date: Mon, 27 Nov 2023 01:38:26 GMT
Accept-Ranges: bytes
Etag: "clea7740a2f00ab5ba65a775e46bc6720678138eb843"
Last-Modified: Tue, 04 Jul 20...



case III: Streamlit



What would you like to say?

Send

你好,小微GPT为您服务,请问有什么可以帮助您?

Tell me how can I secure my MLOps Stack

Securing your MLOps stack is crucial to protect your machine learning operations and data. Here are some steps you can take to enhance the security of your MLOps stack:

1. Secure your infrastructure: Ensure that your infrastructure, including servers, databases, and networking components, is properly secured. This includes implementing firewalls, regularly patching and updating software, and using strong authentication mechanisms.
2. Implement access controls: Control access to your MLOps stack by implementing strong authentication and authorization mechanisms. Use multi-factor authentication (MFA) for user accounts, enforce strong password policies, and limit access privileges to only those who need them.
3. Encrypt data in transit and at rest: Encrypt data when it is being transmitted over networks and when it is stored in databases or file systems. Use secure protocols such as HTTPS for data transmission and implement encryption mechanisms like AES for data at rest.



case III: Streamlit

The screenshot shows a Streamlit application interface. On the left, a sidebar titled "AI client" lists various AI-related functions:

- home page
- AI client**
- identify claims | embeddings | full
- identify claims | embeddings | restricted
- identify claims | function calling | restricted
- identify claims | function calling | full
- identify claims | redact pii
- summarise doc
- use valla tone
- identify claims | human loop recording
- document upload
- nonsense

The main panel has the following sections:

- Describe the AI or system role**: You are a storyteller
- Provide your instructions**: Tell me how to secure my MLOps stack
- Text or article to be assessed (optional)**: (empty text area)
- Select provider and model**: A dropdown menu currently set to "OpenAI | gpt-4-1106-preview". The dropdown list includes:
 - OpenAI | gpt-4-1106-preview
 - OpenAI | gpt-4
 - OpenAI | gpt-3.5-turbo
 - AWS Bedrock | amazon.titan-text-express-v1
 - AWS Bedrock | ai21.j2-mid-v1
 - AWS Bedrock | ai21.j2-ultra-v1
 - AWS Bedrock | anthropic.claude-instant-v1** (highlighted with a red box)
 - ...AWS Bedrock | anthropic.claude-v2, ...warning about potential data breaches, moderation, and inappropriate attacks.

At the bottom of the main panel, a note reads: "The guardians knew that the first line of defense was the mighty Gate of Access Control. Only those who".



- these were some of the most interesting examples of API key leaks/unauthorized access to the models
- now it's time for a bigger picture - **OWASP Top10 projects** and **OWASP AI Exchange**



- **disclaimer:** OWASP Top10 For Web is not the only list of vulnerabilities published by OWASP:
- OWASP Top10 for Web, OWASP Top10 for CI/CD, OWASP Top10 for API, OWASP Top10 for LLMs, OWASP Top10 for ML, OWASP Top10 for Mobile Applications etc.



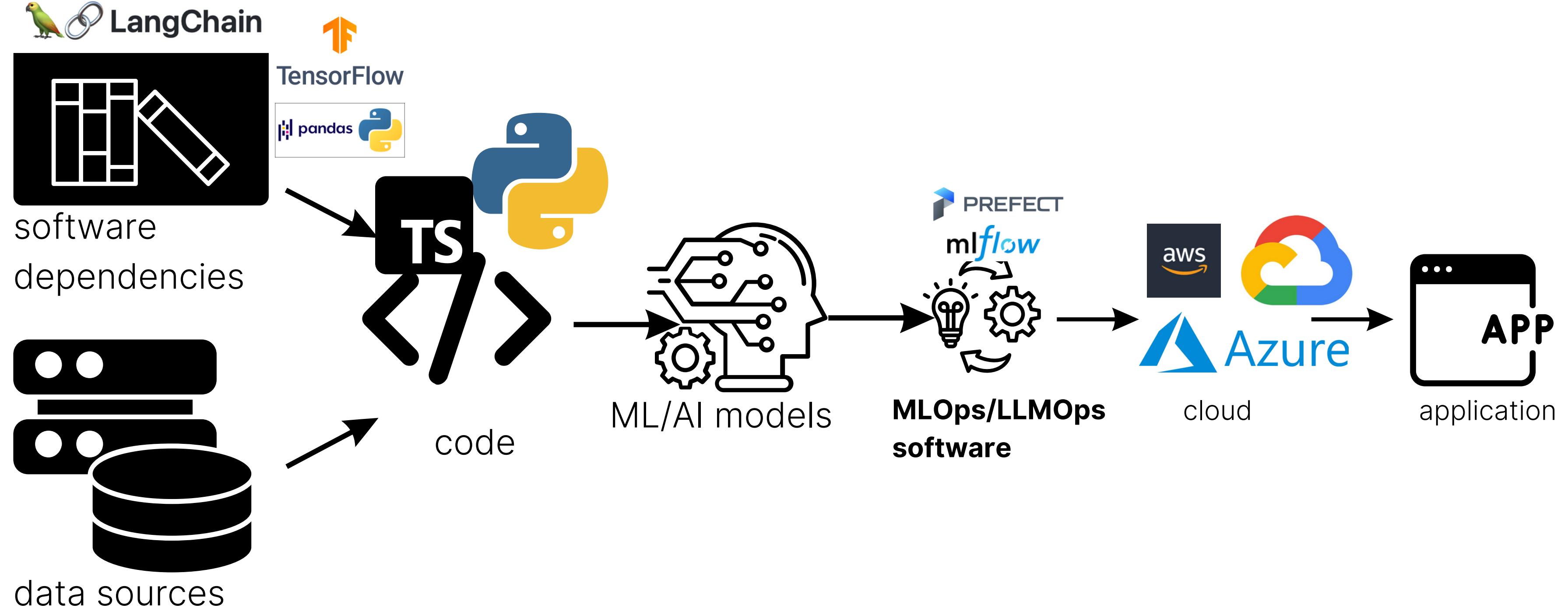
- disclaimer:
of vulnerabilities
- OWASP Top
OWASP Top
OWASP Top
Applications



not the only list
or CI/CD,
or LLMs,
or Mobile



AI Supply Chain



- You can use (at least) the following standards to secure your MLOps stack:
- **OWASP Top10 for Web**
- **OWASP Top10 for CI/CD**
- **OWASP Top10 for API**
- **OWASP Top10 for LLMs OR OWASP Top10 for ML**

OWASP Top10 for ML

- Software Supply Chain security is just one part of the ML security. **OWASP Top10 for ML** tries to cover all of the most important parts of it.

Input Manipulation Attack

Adversarial Attacks, a type of attack in which an attacker deliberately alters input data to mislead the model.

AI Supply Chain Attacks

Data Poisoning Attack

an attacker manipulates the training data to cause the model to behave in an undesirable way.

Transfer Learning Attack

an attacker trains a model on one task and then fine-tunes it on another task to cause it to behave in an undesirable way.

Model Inversion Attack

an attacker reverse-engineers the model to extract information from it.

Model Skewing

an attacker manipulates the distribution of the training data to cause the model to behave in an undesirable way.

Membership Inference Attack

an attacker manipulates the model's training data in order to cause it to behave in a way that exposes sensitive information.

Output Integrity Attack

an attacker aims to modify or manipulate the output of a machine learning model

Model Stealing

an attacker gains access to the model's parameters.

Model Poisoning

an attacker manipulates the model's parameters



OWASP Top10 for ML

- Contributors welcome:

<https://github.com/OWASP/www-project-machine-learning-security-top-10>



OWASP Top10 for LLM

- There's also another - *more specific* - OWASP Top10 list: Top10 for Large Language Models

Prompt Injection	Sensitive Information Disclosure
Insecure Output Handling	Insecure Plugin Design
Training Data Poisoning	Excessive Agency
Model Denial of Service	Overreliance
Supply Chain Vulnerabilities	Model Theft



Check if you can hack LLMs!



source of image: Lakera.AI

<https://gandalf.lakera.ai/>



OWASP AI Exchange

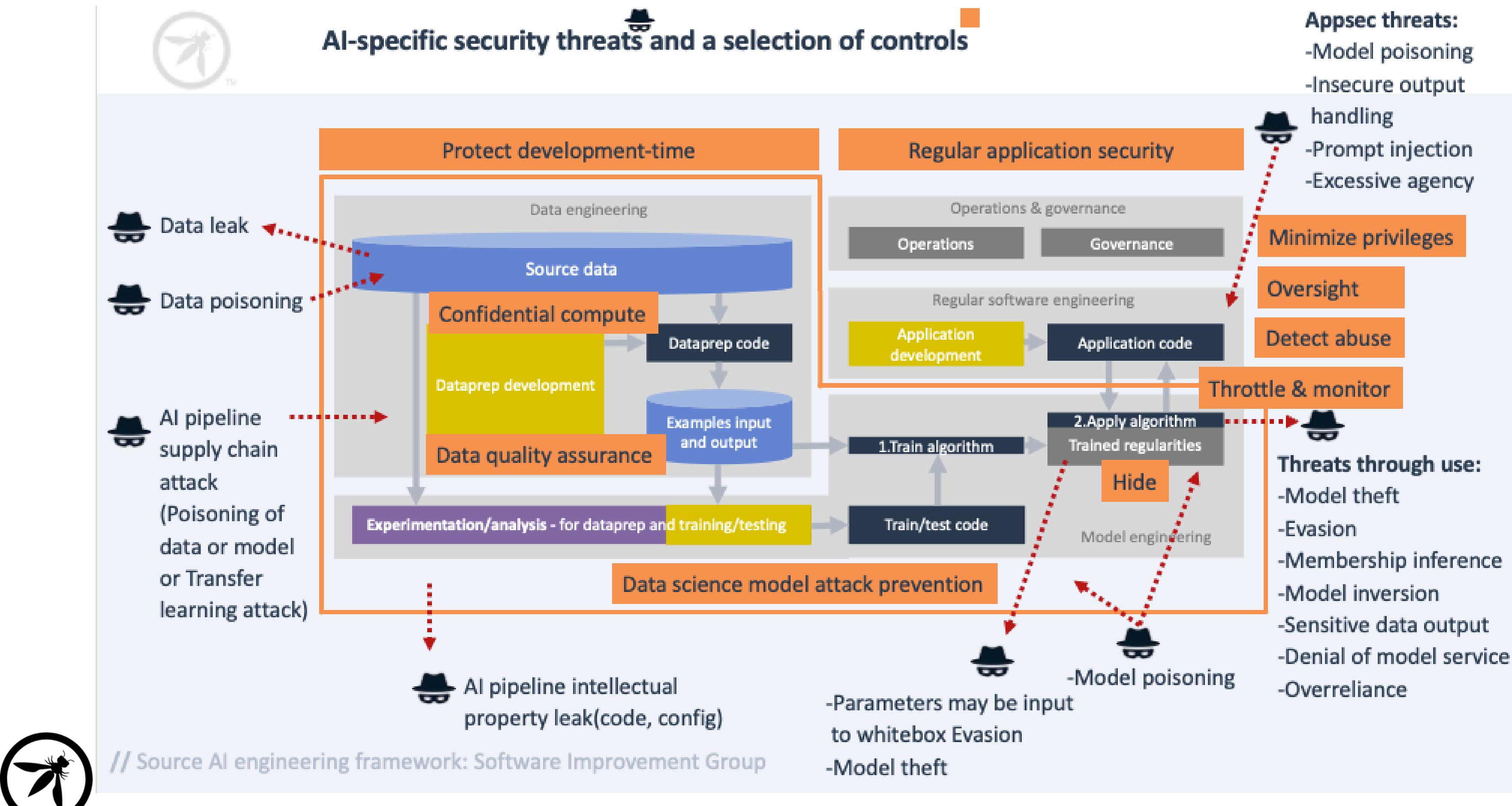
- Document that catalogues the standards and guidelines for AI Security



<https://owasp.org/www-project-ai-security-and-privacy-guide/owaspaiexchange.html>



OWASP AI Exchange



OWASP AI Exchange

A comprehensive overview of AI threats, vulnerabilities, and controls to foster alignment among different standardization initiatives. This includes:

- **EU AI act**
- **ISO/IEC 27090 (AI security)**
- **OWASP ML top 10**
- **OWASP LLM top 10**
- **OpenCRE** - which we want to use to provide the AI Exchange content through security chatbot **OpenCRE-Chat**.



Key takeaways

- Don't trust OSS providers blindly
- Remember that when you deploy self-hosted instance of an app, you're the one responsible for the security
- Use OWASP Top10s as a starting point for securing your applications, and then proceed with more detailed standards/guidelines
- Keep in mind that when you secure ML/LLM apps, you need to face both “classical” and novel threats for your apps
- If you don't know how to make your LLMOps secure, just ask your LLM ;)

