



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**



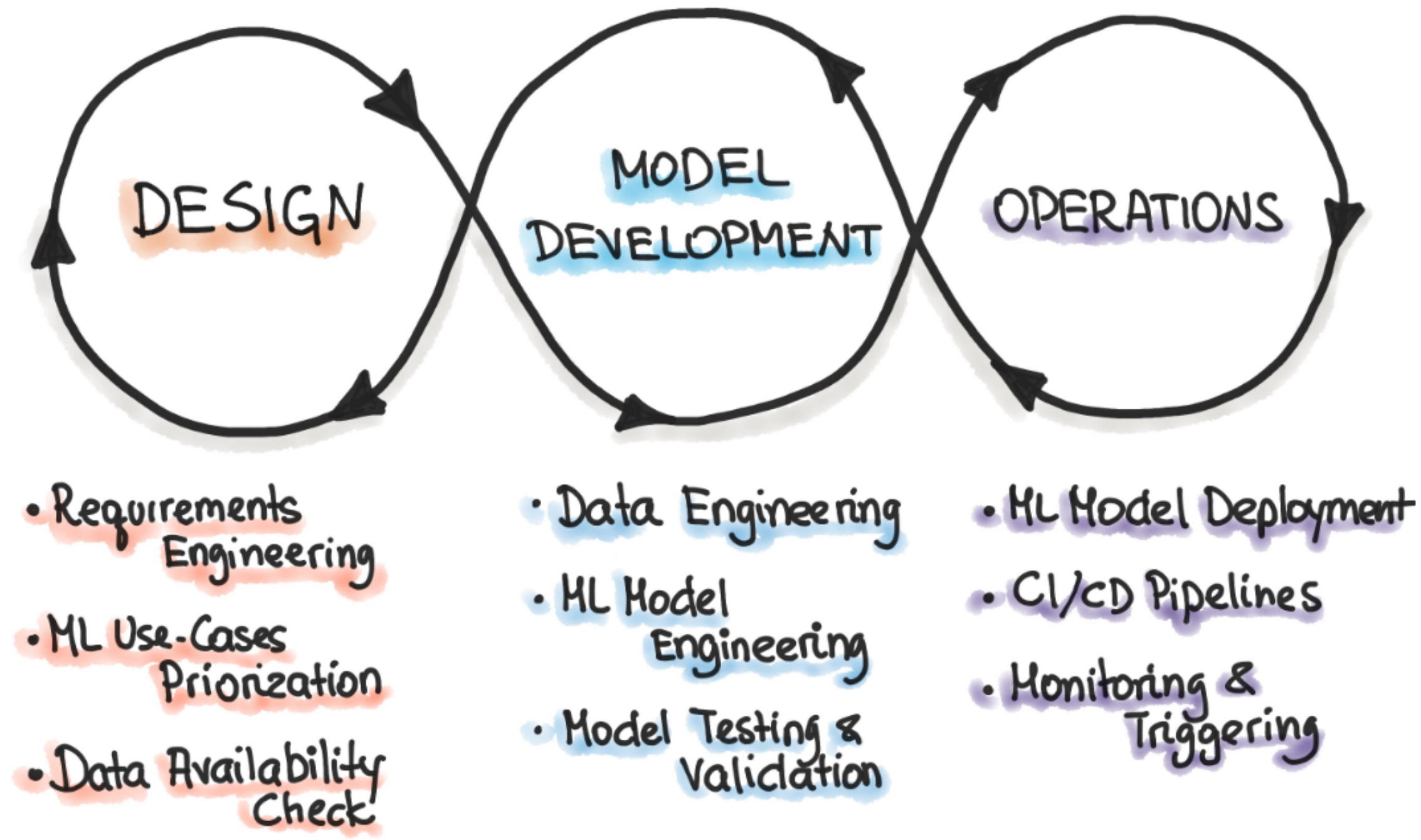
What will we talk about?

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



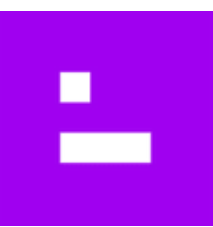
MLOps

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



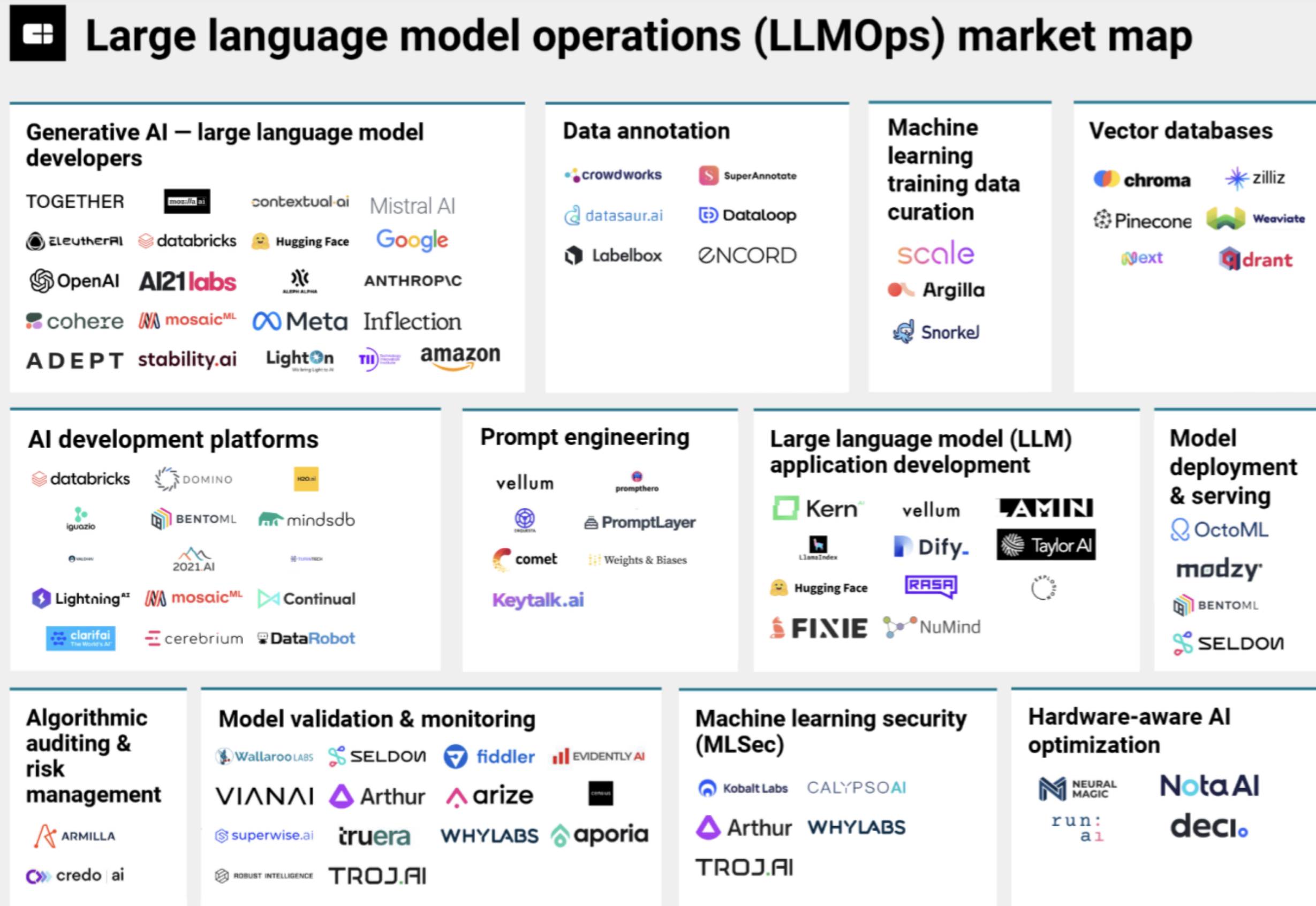
Model deployment



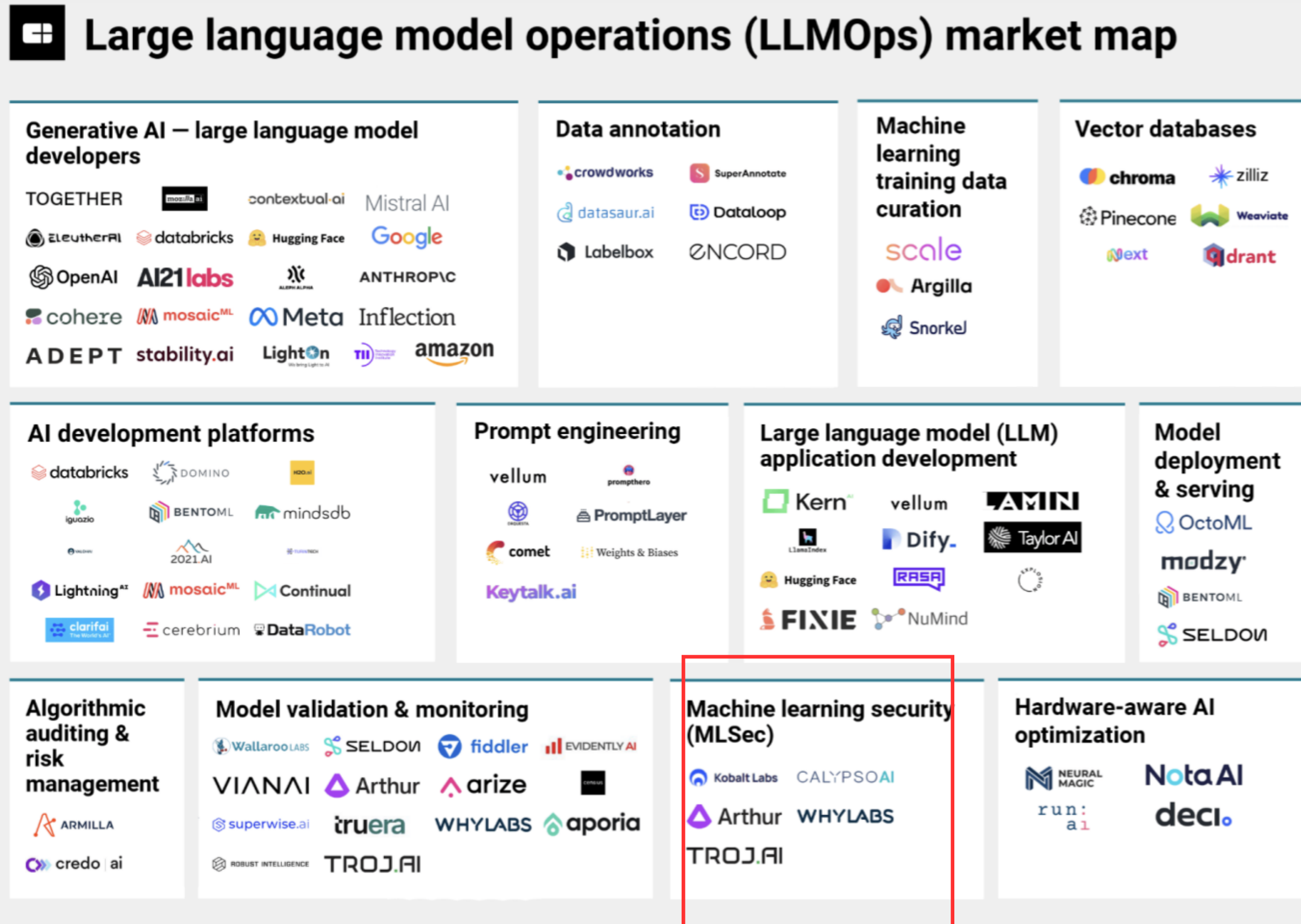
and many more...



There's LLMOps market as well:



There's LLMOps market as well:



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

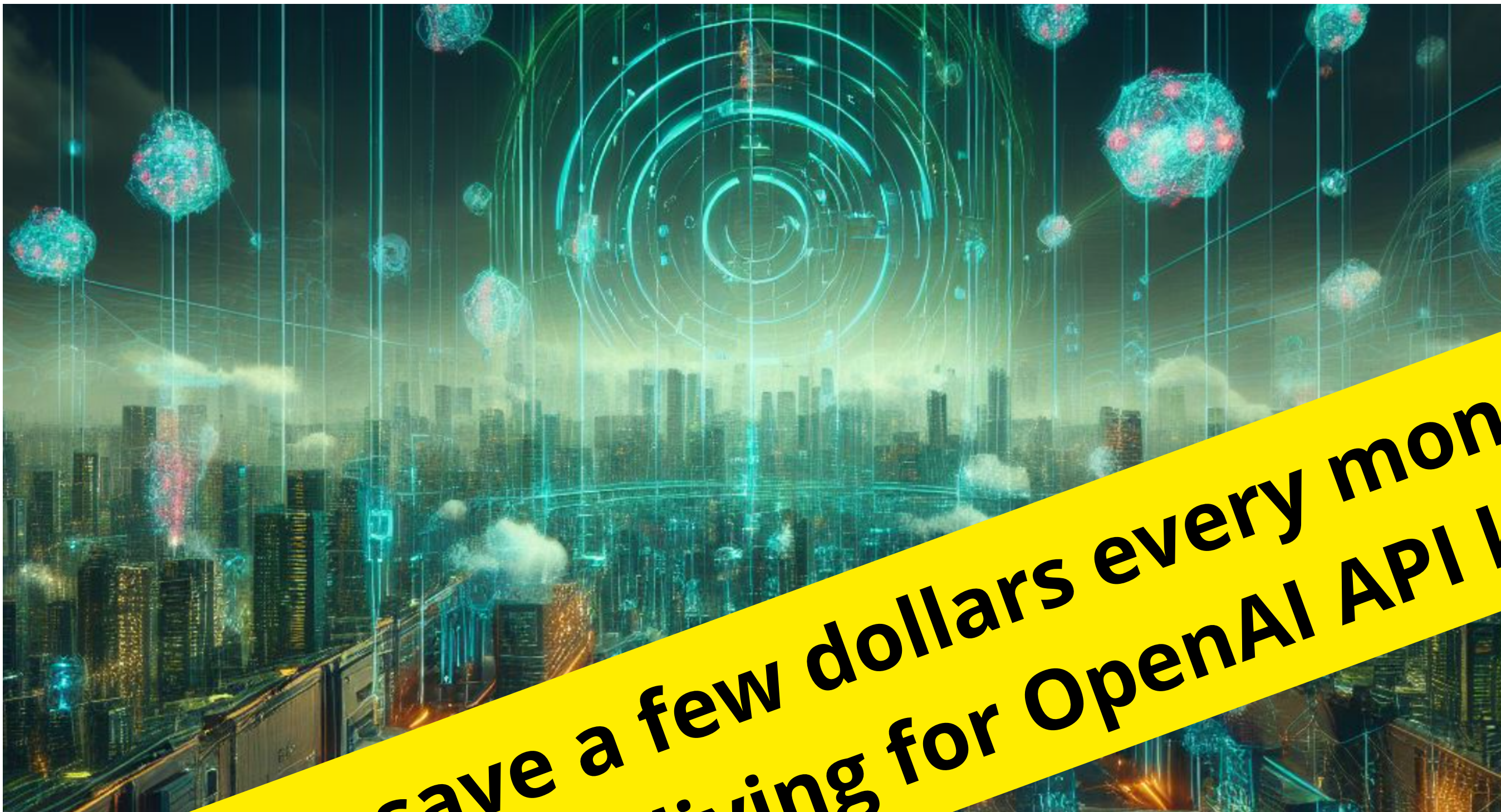


title update:



~~The ML(Ops) Security Landscape~~





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

~~Security Landscape~~



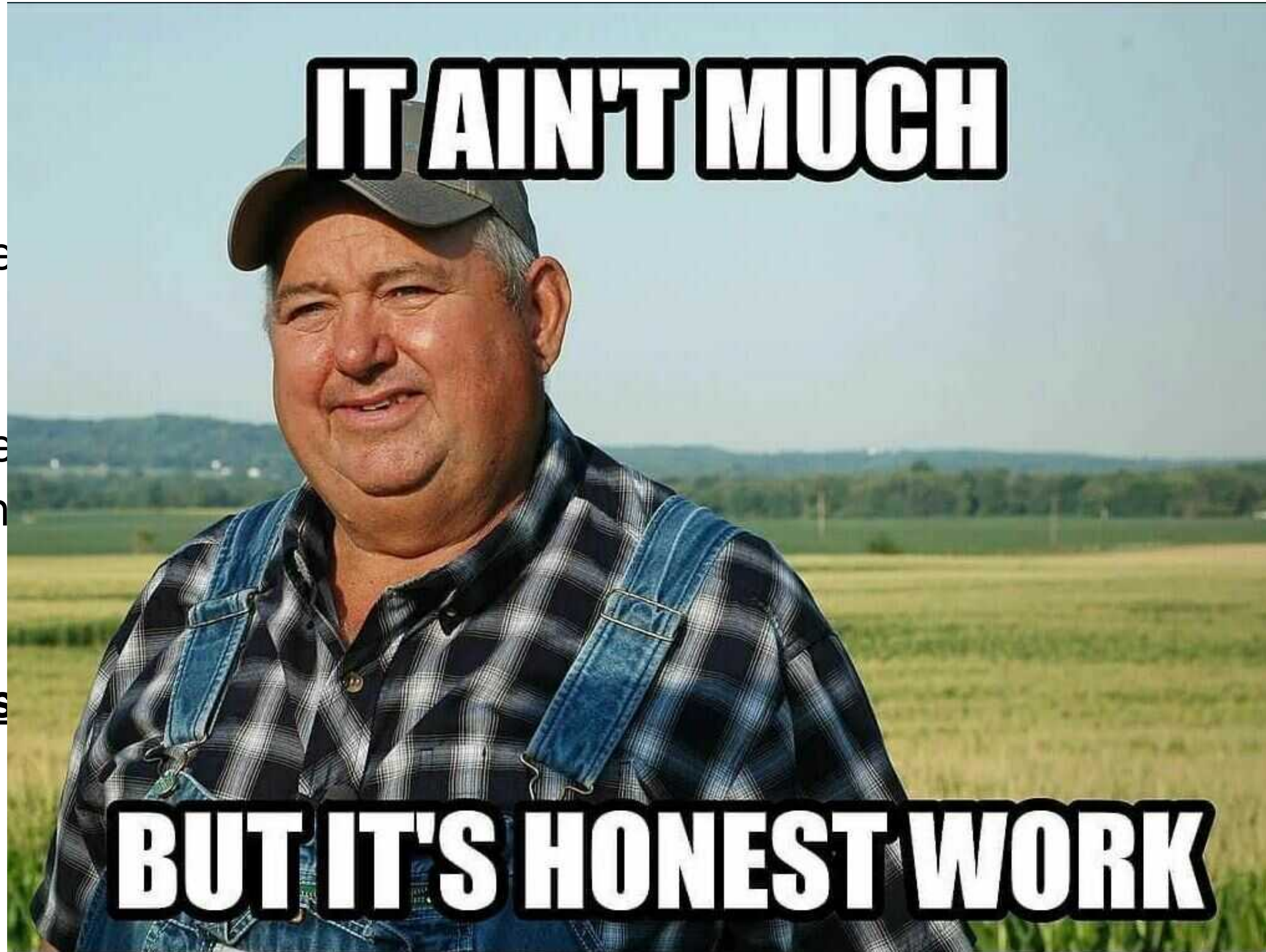
my approach

- 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 data is
- I wanted miscon
- Due to MLOps



f some

ne

esome



What is Shodan

Google

surface web

SHODAN
censys

deep web

GORCH

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...](#)

[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
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

HTTP/1.1 200 OK
Date: Sun, 26 Nov 2023
Server: Apache/2.4.46
X-Powered-By: PHP/5.6.
Set-Cookie: OCSESSID=b
Set-Cookie: OCSESSID=0
Set-Cookie: language=e

IPEE Buddy - smart toilet assistant - IPEE [↗](#)

164.92.151.165
www.buddy.ipee.eu
home.ipee.eu
www.home.ipee.eu
ipee.eu
about.ipee.eu
DigitalOcean, LLC
Netherlands, Amsterdam



cloud

SSL Certificate

Issued By:
|- Common Name:
R3

|- Organization:
Let's Encrypt

Issued To:
|- Common Name:
ipee.eu

Supported SSL Versions:
TLSv1.2, TLSv1.3

HTTP/1.1 200 OK
Date: Sun, 26 Nov 2023
Server: Apache/2.4.41
Link: <https://ipee.eu
Vary: Accept-Encoding
Cache-Control: no-cach
Pragma: no-cache
Expires: 0
Access-Control-Allow-0
Transfer-Encod...



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"
```

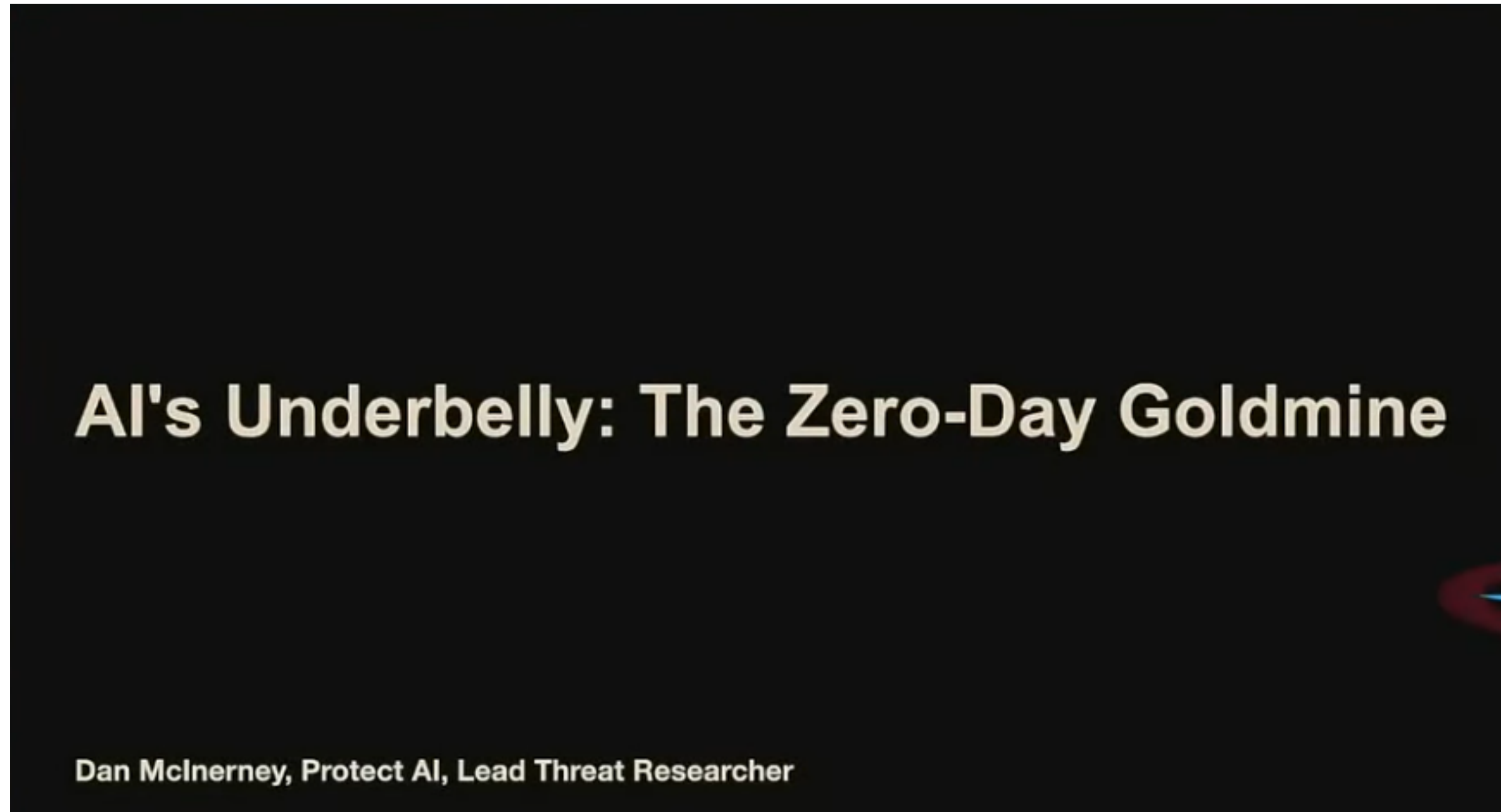
```
http.html:"Prefect Server"
```

```
.....
```



Kudos for inspiration

- Dan McInerney from ProtectAI @ ROOTCON



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



Case I: Flowise & Langflow

Simple LLM Chain

OpenAI

Inputs

OpenAI Api Key *

Model Name

text-davinci-003

Temperature

0.7

Output

OpenAI

Prompt Template

Inputs

Template *

What is a good name for a company that makes {product}?

Output

PromptTemplate

LLM Chain

Inputs

LLM *

Prompt *

Format Prompt Values

```
{  
  "input_language": "English",  
  "output_language": "French"  
}
```

Output

LLMChain

Company that makes t shirt

Tee Time Apparel.

Colorful socks

Rainbow Socks Co.

Motorbikes

MotoRiders.

Type your question...



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

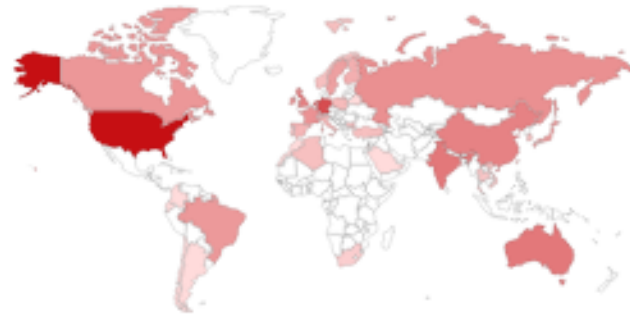


http.title:"flowise"

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

- Organization:

Let's Encrypt

Issued To:

- Common Name:

n.i-setka.ru

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

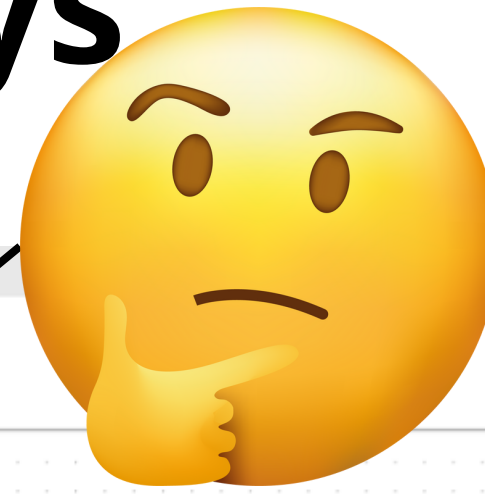


/api/v1/database/export

```
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: h[REDACTED]
12 Accept-Encoding: gzip, deflate, br
13 Accept-Language: en-US,en;q=0.9
14 Connection: close
15
16
```

```
1 {
2   "id": "bf[REDACTED]",
3   "role": "userMessage",
4   "chatflowid": "[REDACTED]",
5   "content": "Bonjour [REDACTED]",
6   "sourceDocuments": null,
7   "createdAt": [REDACTED]
8 },
9 {
10  "id": "76[REDACTED]",
11  "role": "apiMessage",
12  "chatflowid": "e[REDACTED]",
13  "content":
14    "Bien [REDACTED]
15    us aujourd'hui [REDACTED]",
16  "sourceDocuments": null,
17  "createdAt": [REDACTED]
18 },
19 {
20  "id": "[REDACTED]",
21  "role": "userMessage",
22  "chatflowid": "e[REDACTED]",
23  "content": "déjà, [REDACTED] ? ",
24  "sourceDocuments": null,
25  "createdAt": "2023-09-07T09:49:47.000Z"
26 },
27 {
28  "id": "7[REDACTED]",
29  "role": "apiMessage",
30  "chatflowid": "[REDACTED]",
31  "content":
32    "Je sui [REDACTED], l
33    Legani [REDACTED] ses
34    [REDACTED] oni
35    [REDACTED] comment puis-je vous aider
36  "sourceDocuments": null,
37  "createdAt": "[REDACTED]"
38 }
39 }
```

Flowise - leaking API keys



The screenshot shows a Flowise workflow titled "Simple LLM Chain". It consists of three nodes: "OpenAI", "Prompt Template", and "LLM Chain".

- OpenAI Node:** Contains an "Inputs" section with "OpenAI Api Key*" (highlighted with a red box and an arrow from the emoji), "Model Name" (text-davinci-003), and "Temperature" (0.7). The "Output" section is labeled "OpenAI".
- Prompt Template Node:** Contains a "Template*" field with the text "What is a good name for a company that makes {product}?". The "Output" section is labeled "PromptTemplate".
- LLM Chain Node:** Contains "Inputs" for "LLM*" and "Prompt*", and a "Format Prompt Values" field with a JSON object:

```
{ "input_language": "English", "output_language": "French" }
```

. The "Output" section is labeled "LLMChain".

Arrows indicate the flow of data from the OpenAI node to the Prompt Template node, and from the Prompt Template node to the LLM Chain node. A red box highlights the "OpenAI Api Key*" field, with a black arrow pointing from a thinking face emoji above it.




On the right side, there is a chat window showing a list of suggestions for a company that makes t-shirts:

- Company that makes t shirt
- Tee Time Apparel.
- Colorful socks
- Rainbow Socks Co.
- Motorbikes
- MotoRiders.

At the bottom of the chat window is a text input field with the placeholder "Type your question..." and a send button.

This example comes from the Flowise docs

Flowise - leaking API keys

 ChatOpenAI  

Inputs

OpenAI Api Key *




Model Name
gpt-3.5-turbo

Temperature
0.9

[Additional Parameters](#)

Output

ChatOpenAI

 Github  

Inputs

Text Splitter

Repo Link *
github.com/[REDACTED]

Branch *
main

Access Token

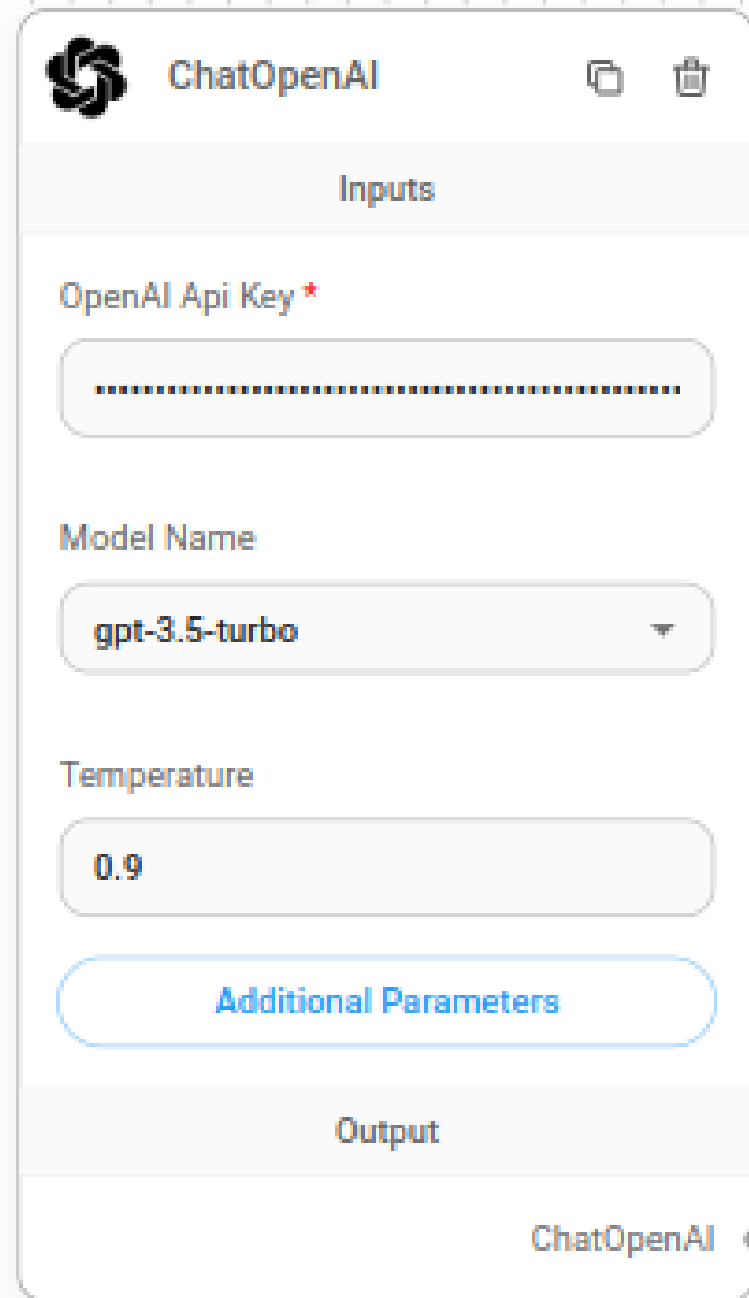
[Additional Parameters](#)

Output

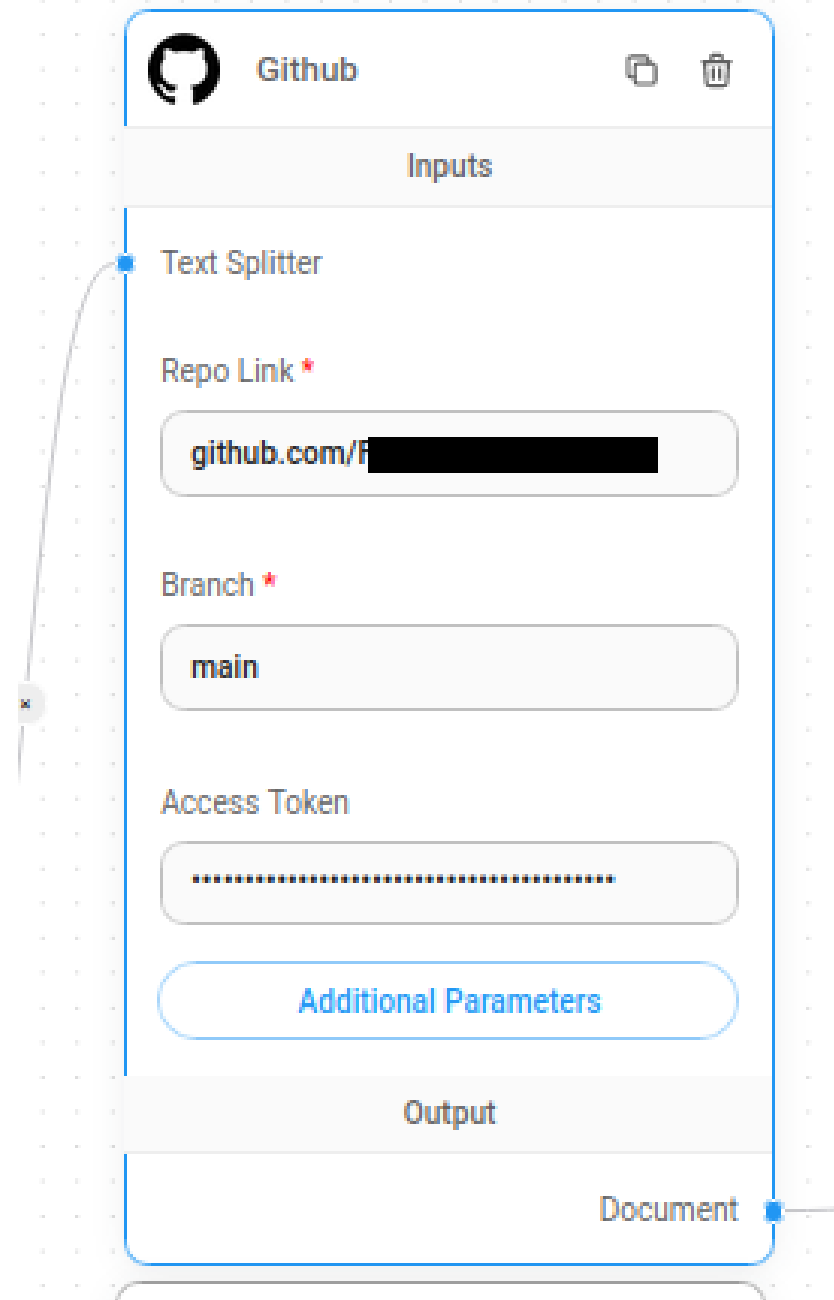
Document

Flowise - leaking API keys

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



The screenshot shows the configuration interface for the 'ChatOpenAI' node. It features a header with the OpenAI logo and the text 'ChatOpenAI'. Below this is an 'Inputs' section containing three fields: 'OpenAI Api Key *' (a redacted text input), 'Model Name' (a dropdown menu with 'gpt-3.5-turbo' selected), and 'Temperature' (a text input with '0.9'). There is also an 'Additional Parameters' button. At the bottom, there is an 'Output' section with the label 'ChatOpenAI'.



The screenshot shows the configuration interface for the 'Github' node. It features a header with the GitHub logo and the text 'Github'. Below this is an 'Inputs' section containing four fields: 'Text Splitter' (a blue dot), 'Repo Link *' (a text input with 'github.com/' followed by a redacted name), 'Branch *' (a text input with 'main'), and 'Access Token' (a redacted text input). There is also an 'Additional Parameters' button. At the bottom, there is an 'Output' section with the label 'Document'.

Flowise - leaking API keys

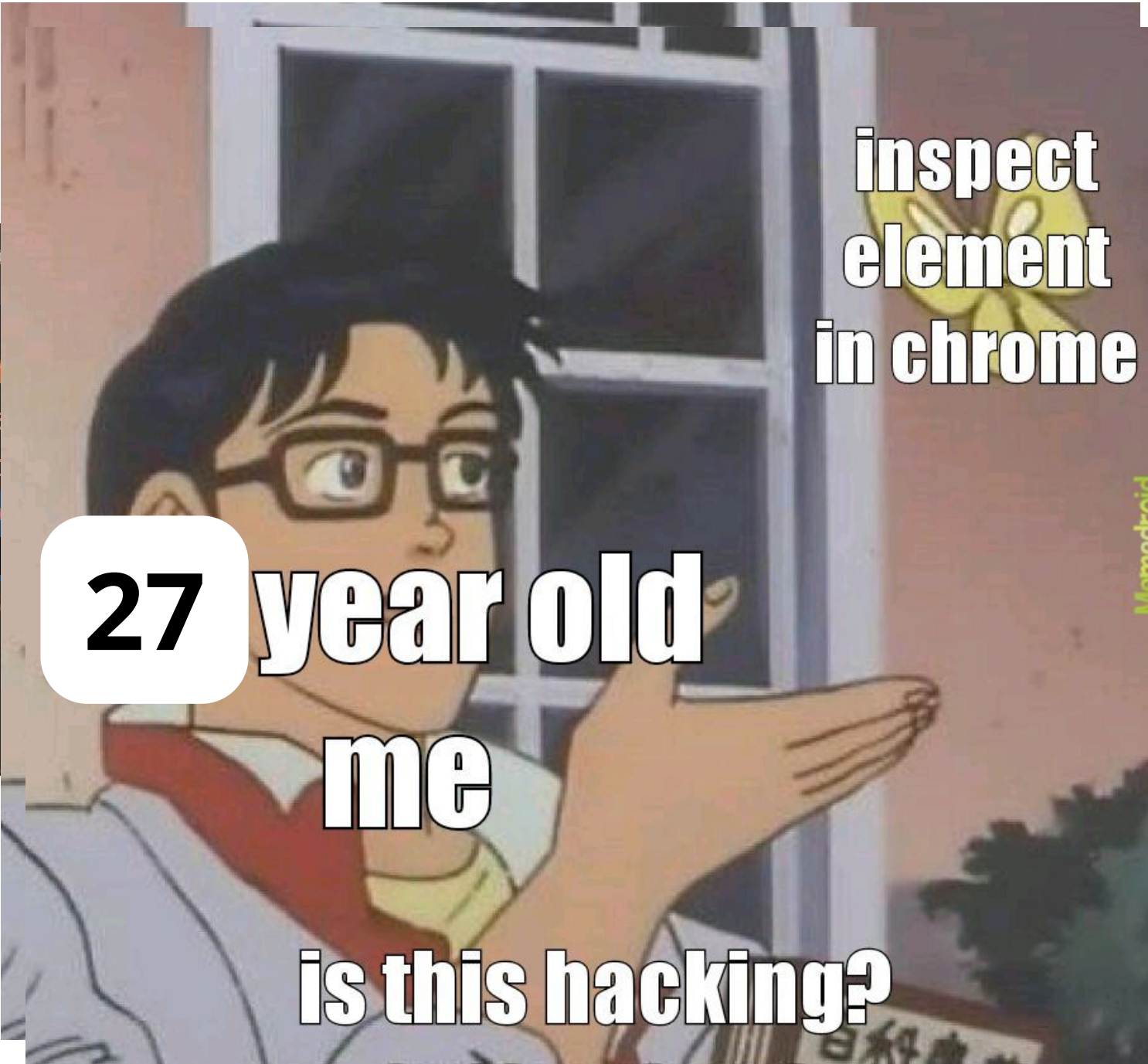
If we just inspect an element in the browser it would be too easy, right? **No**

```
Recorder  Console Sources Network Performance Memory Application Security Lighthouse DOM Invader
```

```
▼ <div class="MuiBox-root css-19midj6">  
  ▶ <div style="display: flex; flex-direction: row;"> ... </div> flex  
  ▼ <div class="MuiFormControl-root css-7e6n2m"> flex  
    ▼ <div class="MuiInputBase-root MuiOutlinedInput-root MuiInputBase-colorPrimary MuiInputBase-formControl MuiInputBase-sizeSmall css-9xlxsh"> flex  
      <input aria-invalid="false" id="openAIApiKey" name="openAIApiKey" rows="1" type="password" step="0.1" class="MuiInputBase-input MuiOutlinedInputBase-input MuiInputBase-sizeSmall css-gs47q4" value="sk-KnP[REDACTED]" style="height: inherit;"> == $0  
      ▶ <fieldset aria-hidden="true" class="MuiOutlinedInput-notchedOutline css-ncxurb"> ... </fieldset>  
    </div>  
  </div>  
</div>  
</div>
```

Flowise - leaking API keys

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



```
DOM Invader  
control MuiInputBase-sizeSmall css-9xlxsh"> flex  
p="0.1" class="MuiInputBase-input MuiOutlinedIn  
inherit;"> == $0
```


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

Hey Mikołaj,

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.

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

flowise not allowing it:

Hey Mil

Thanks

Curious

Rhar I li



them as Flowise doesn't allow it.

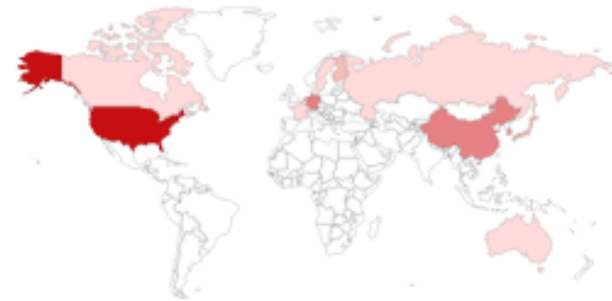
yl

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](#)

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: uvicorn
content-type: text/html; charset=utf-8
content-length: 724
last-modified: Tue, 25 Jul 2023 00:13:54 GMT
etag: 46ff59872e5a8cd55db1ec1cf25dbaf5
```

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)

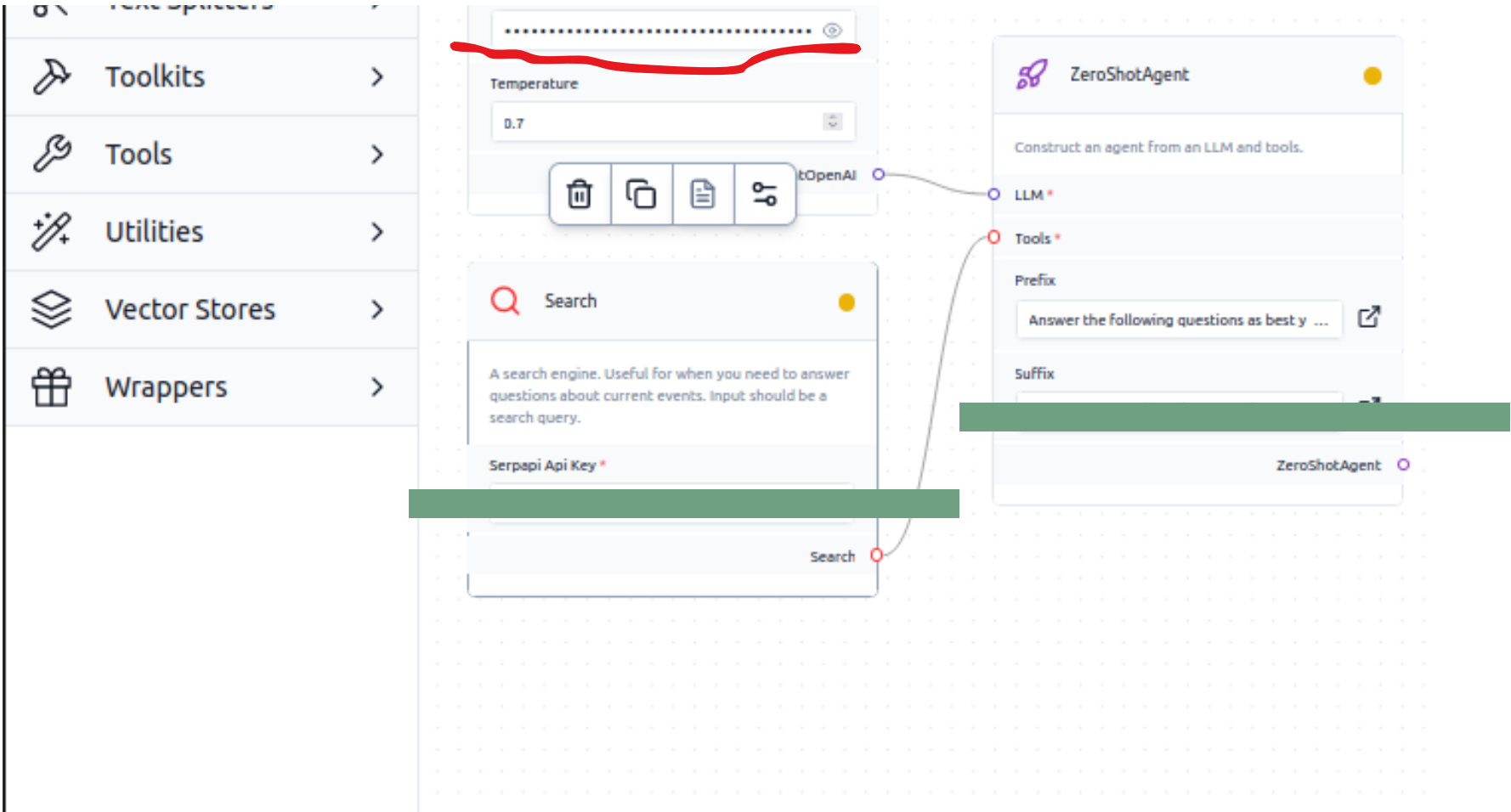
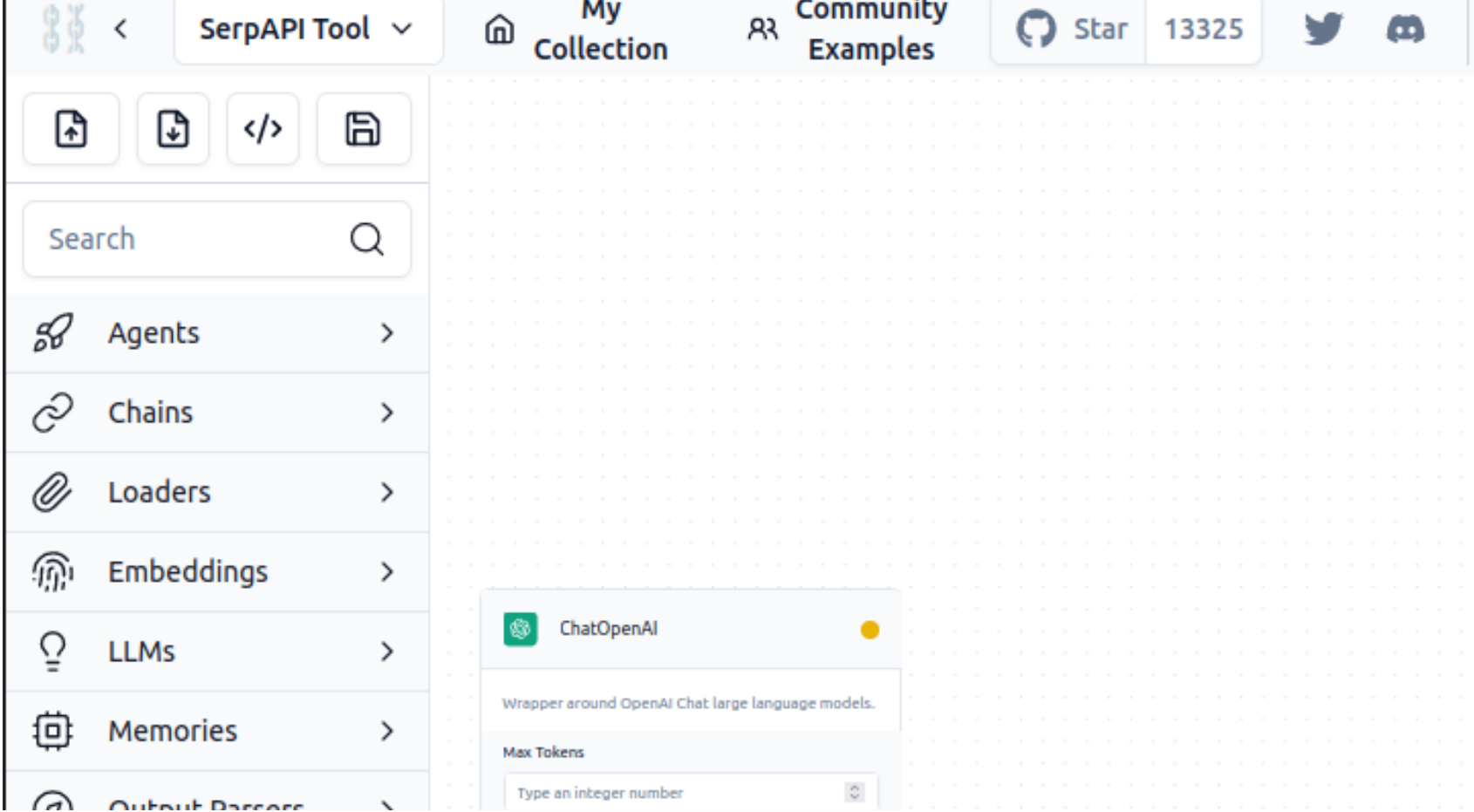
The screenshot displays the Langflow web interface. On the left is a sidebar menu with categories: Agents, Chains, Loaders, Embeddings, LLMs, Memories, Output Parsers, Prompts, Retrievers, Text Splitters, Toolkits, Tools, Utilities, Vector Stores, and Wrappers. The main workspace shows a workflow on a grid background. It includes a 'ChatOpenAI' component with fields for Max Tokens, Model Name (gpt-3.5-turbo-0613), OpenAI API Base, OpenAI API Key (highlighted with a red circle), and Temperature (0.7). Below it is a 'Search' component with a description and a 'Serpapi Api Key' field (highlighted with a green bar). To the right is a 'ZeroShotAgent' component with fields for LLM, Tools, Prefix, and Suffix. The workflow is connected by lines, and a 'ZeroShotAgent' component is also 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 eyxxxxxxxxxmU
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-6xxxxxxxxxxxxx0R",
"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

TOTAL RESULTS
281

TOP COUNTRIES



United States	169
Korea, Republic of	38
United Kingdom	21
France	20
Germany	9
More...	

TOP PORTS

80	122
3000	83
443	46
5000	5
81	3
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 Shodan account upgrade options.](#)

BentoML Prediction Service

35.172.210.62
ec2-35-172-210-62.compute-1.
amazonaws.com

[Amazon Technologies Inc.](#)

United States, Ashburn



cloud

HTTP/1.1 200 OK
Date: Mon, 27 Nov 2023 00:14:01 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 2859
Connection: keep-alive
server: unicorn

BentoML Prediction Service

18.234.8.154
ec2-18-234-8-154.compute-1.a
mazonaws.com

[Amazon Technologies Inc.](#)

United States, Ashburn



cloud

HTTP/1.1 200 OK
Date: Sun, 26 Nov 2023 23:33:49 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 2859
Connection: keep-alive
server: unicorn

BentoML Prediction Service

3.123.234.211
ec2-3-123-234-211.eu-central-
1.compute.amazonaws.com

[A100 ROW GmbH](#)

Germany, Frankfurt am
Main



cloud

HTTP/1.1 200 OK
date: Sun, 26 Nov 2023 23:26:03 GMT
server: unicorn
content-length: 2945
content-type: text/html; charset=utf-8



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": "meta-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

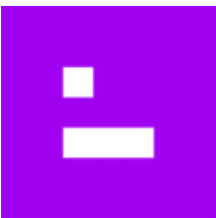
Code Details

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 users to the MLOps tools.\n3. Use encryption to protect sensitive data stored in the MLOps tools.\n4. Regularly update the MLOps tools to address any known vulnerabilities.\n5. Conduct regular security audits to identify and address any potential security vulnerabilities.\n6. Develop an incident response plan to quickly respond to any security incidents.\n7. Train your employees on the importance of security and data protection.\n8. Use a secure network infrastructure to prevent unauthorized access to the MLOps tools.\n9. Implement a robust backup strategy to ensure that the MLOps tools are protected from external threats.\n10. Implement a robust data governance strategy to ensure that the MLOps tools are managed securely.\n\nBy implementing these security measures, you can secure your MLOps tools and protect your organization from potential security threats."
    }
  ]
}
```

1. Implement a strong authentication mechanism to ensure that only authorized users can access the MLOps tools.



case III

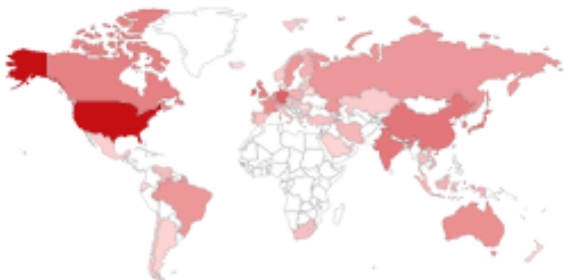
- but what if you wanted to get the output that would be easier to read? And maybe you wanted to compare multiple LLMs in one place? **Misconfigured MLOps tools got you covered!**



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](#)

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: "c1ea7740a2f00ab5ba65a775e46bc6720678138eb843"
Last-Modified: Wed, 12 Jul 20...

Streamlit [↗](#)

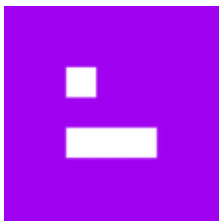
3.39.191.164
ec2-3-39-191-164.ap-northeast-2.compute.amazonaws.com

[AWS Asia Pacific \(Seoul\) Region](#)

Korea, Republic of, Seoul

cloud

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: "c1ea7740a2f00ab5ba65a775e46bc6720678138eb843"
Last-Modified: Tue, 04 Jul 20...



case III: Streamlit

 XW GPT

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 for configuring an AI client. On the left is a sidebar menu with the following items: home page, AI client (highlighted), 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, and nonsense. The main content area is titled 'Describe the AI or system role' and contains a text input field with the text 'You are a storyteller'. Below this is a section 'Provide your instructions' with a text input field containing 'Tell me how to secure my MLOps stack'. The next section is 'Text or article to be assessed (optional)' with an empty text input field. The final section is 'Select provider and model', which features a dropdown menu. The dropdown is open, showing a list of options: 'OpenAI | gpt-4-1106-preview' (selected), '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', and 'AWS Bedrock | anthropic.claude-instant-v1'. A red hand-drawn circle highlights the dropdown menu and its options. At the bottom of the page, there is a snippet of text: '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**

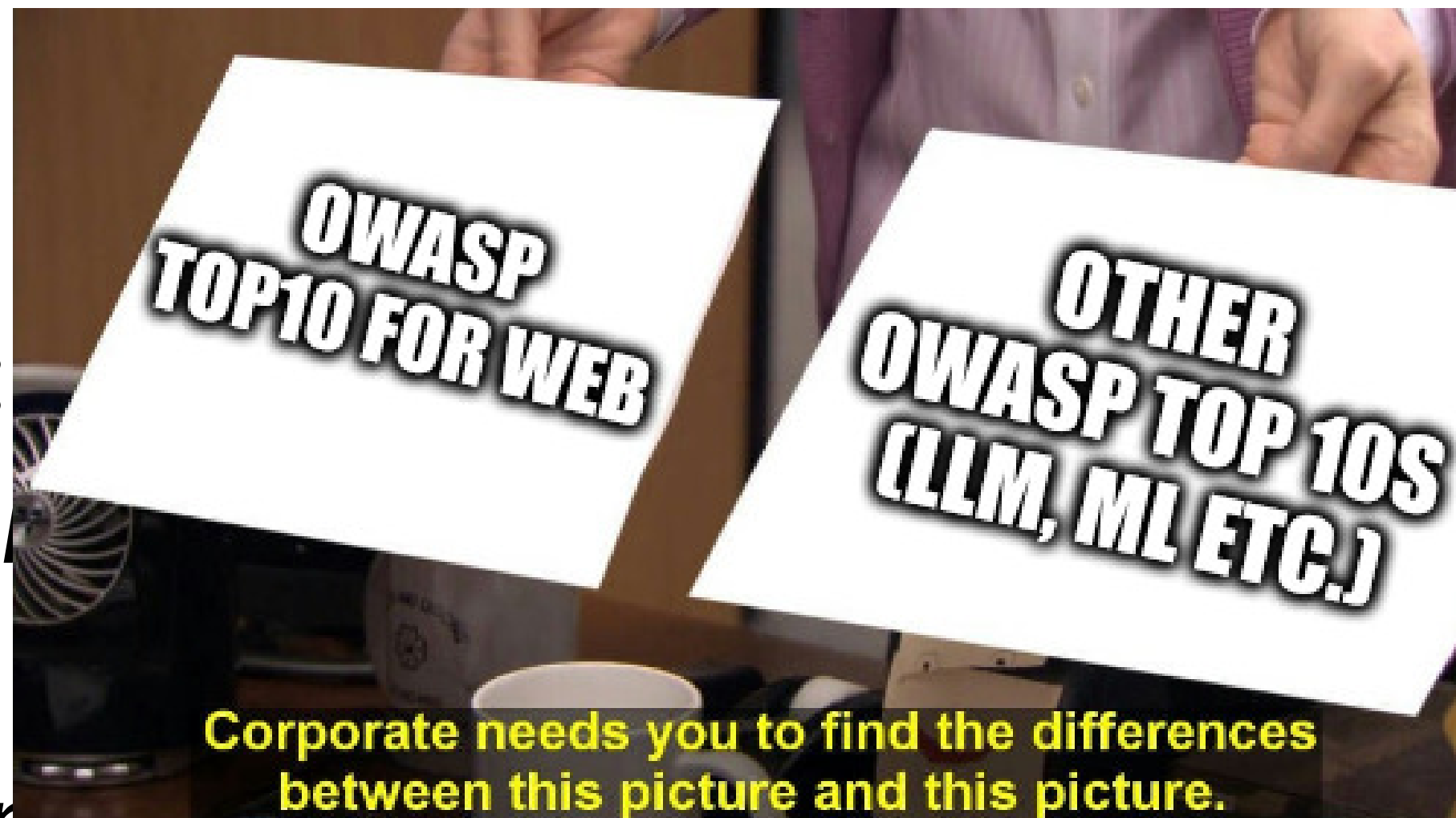


- **disclaimer:** *OWASP Top10 For Web is not the only list of vulnerabilities published by OWASP:*
- *OWASP Top10 for Web, OWASP Top10 for CI/CD, OWASP Top10 for API, OWASP Top10 for LLMs, OWASP Top10 for ML, OWASP Top10 for Mobile Applications etc.*



- **disclaimer:**
of vulnerability

- OWASP Top 10 for Web Applications
- OWASP Top 10 for API Applications
- OWASP Top 10 for Mobile Applications



not the only list

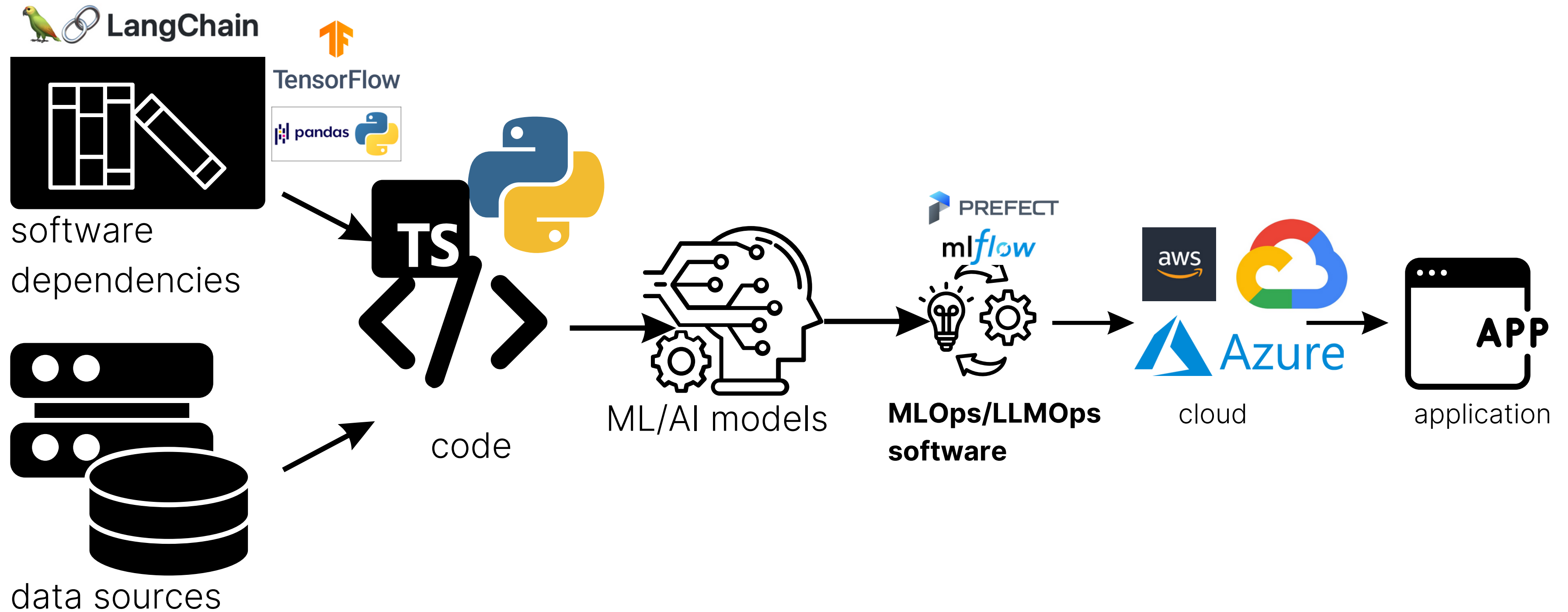
for CI/CD,

for LLMs,

for 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!



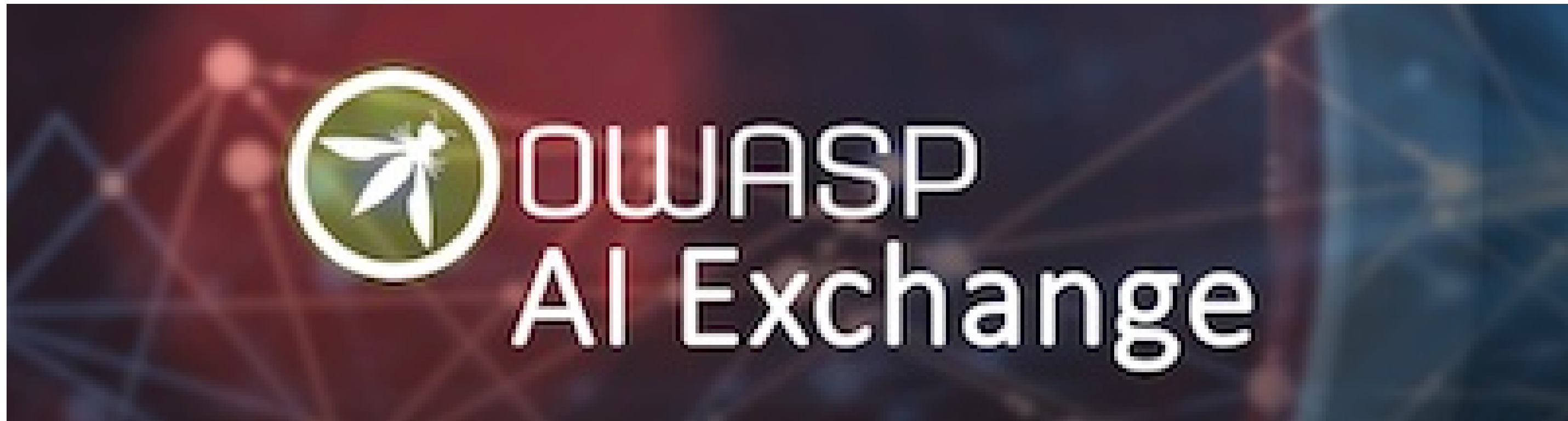
source of image: Lakera.AI

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



OWASP AI Exchange

- Document that catalogues the standards and guidelines for AI Security



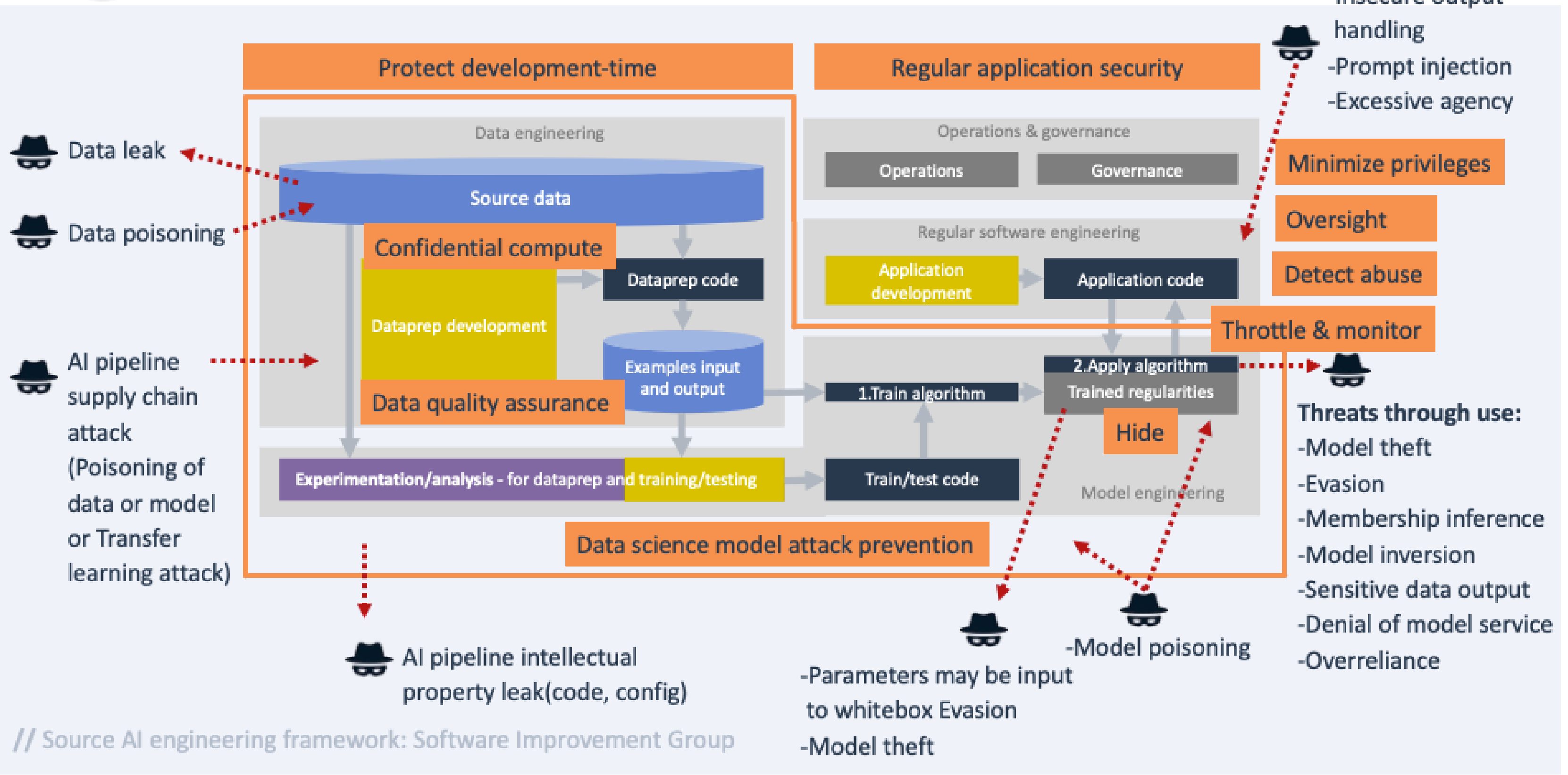
<https://owasp.org/www-project-ai-security-and-privacy-guide/owaspaiexchange.html>



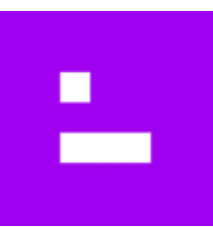
OWASP AI Exchange



AI-specific security threats and a selection of controls



// Source AI engineering framework: Software Improvement Group



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 ;)

