

The Language of Security

Exploring Practical Applications of LLMs in Cyber Security

Jesse van der Zweep



Table of contents

01

How to think about LLMs

03

Challenges and
opportunities

02

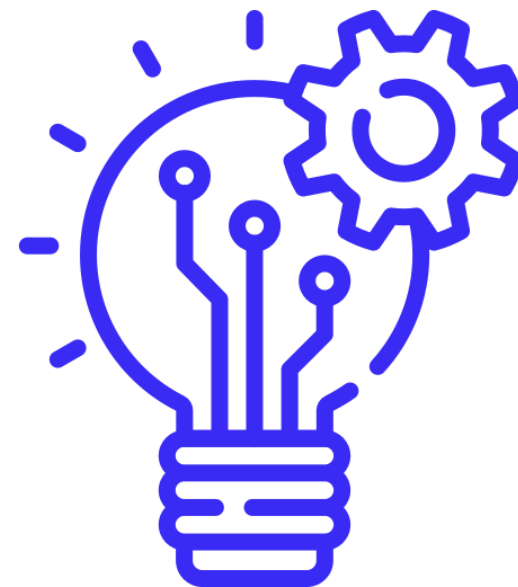
Demos

04

Q&A

01

How to think about LLMs



They predict the next word

I hope this email finds you well. I am writing to

Given this
context

They predict the next word

I hope this email finds you well. I am writing to extend

Given this
context

Predict the
next word

They predict the next word

I hope this email finds you well. I am writing to

...

offer

discuss

extend

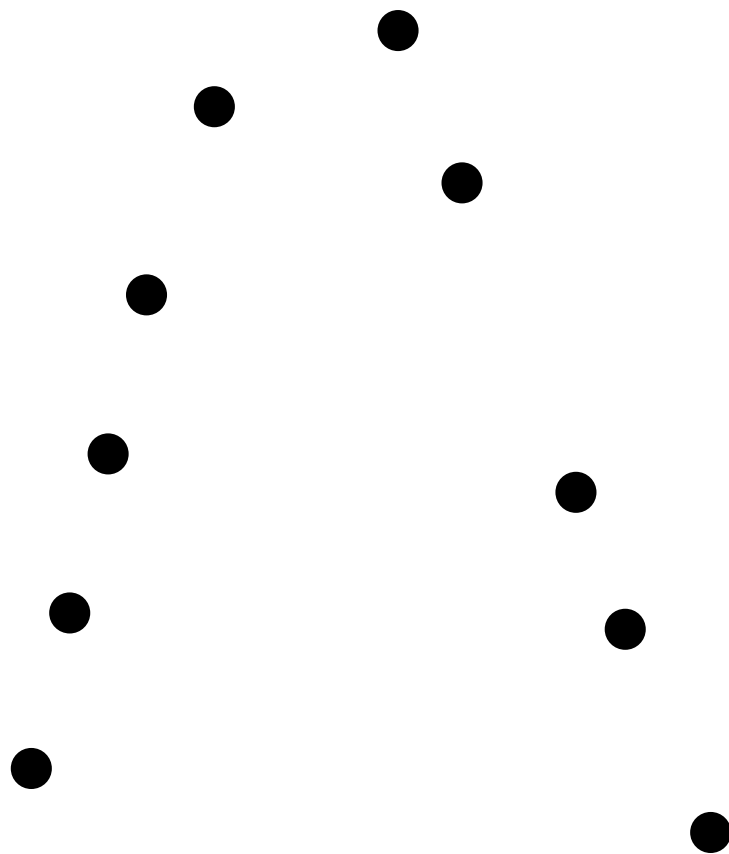
share

request

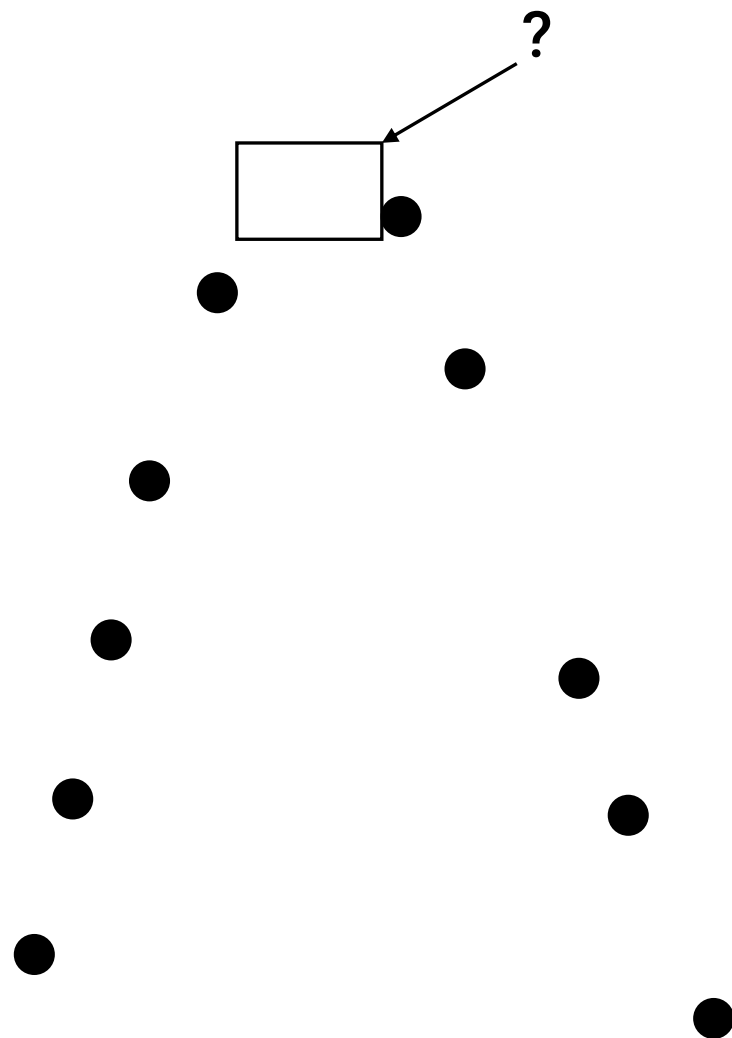
express

...

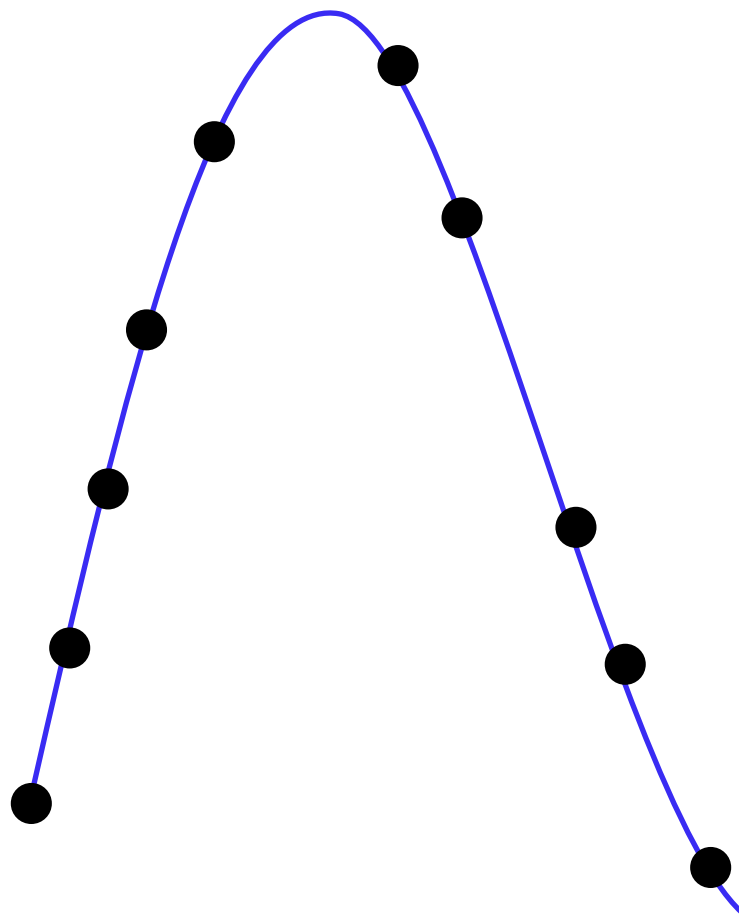
They can
interpolate



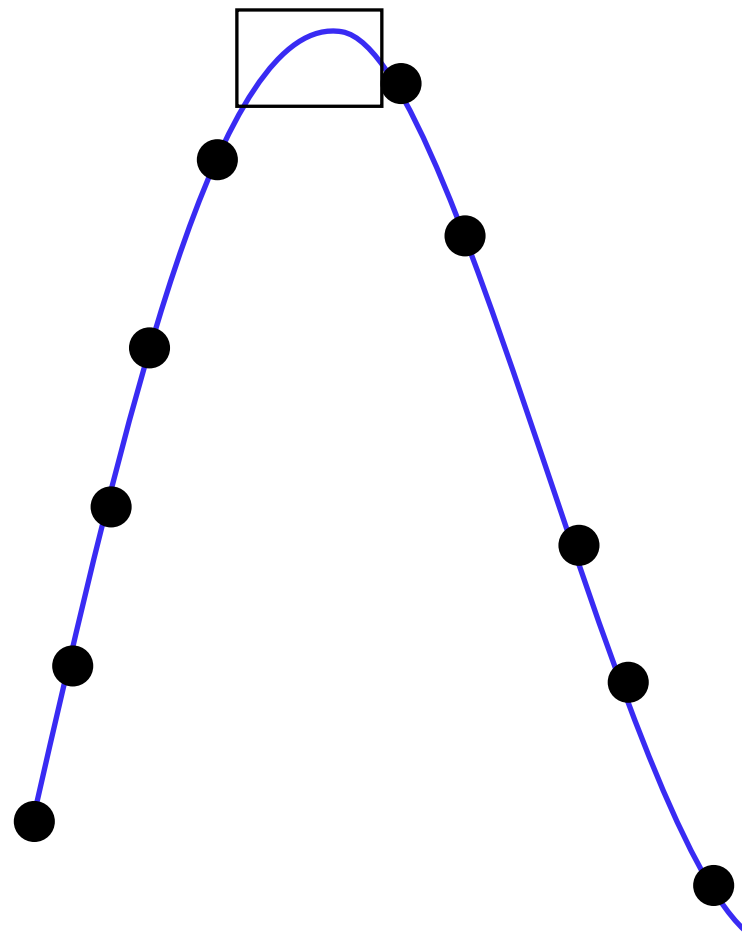
They can
interpolate



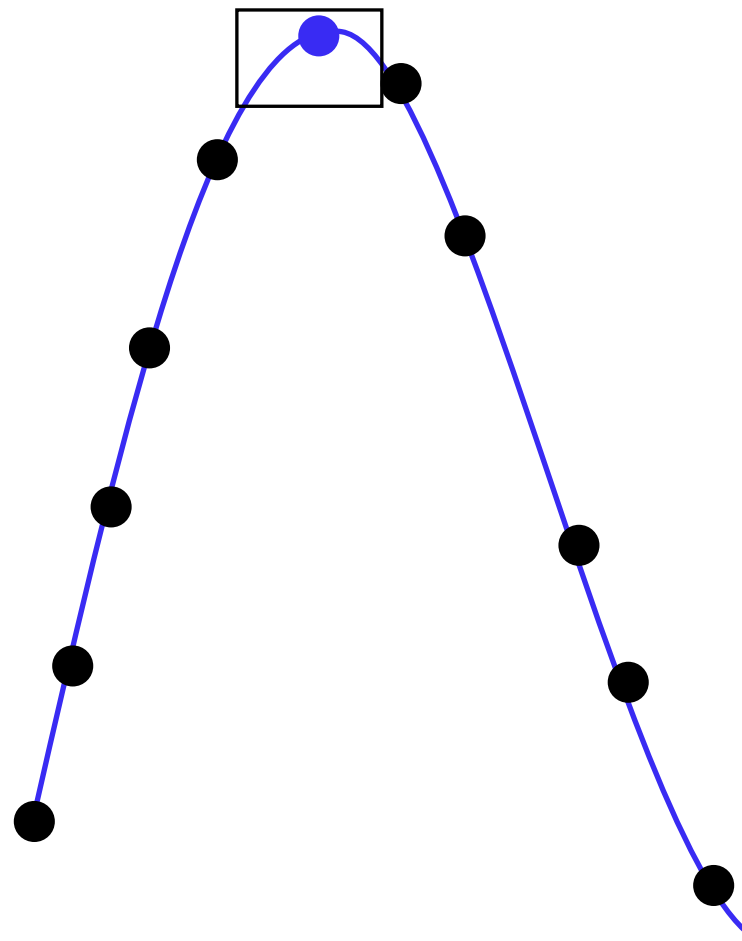
They can
interpolate



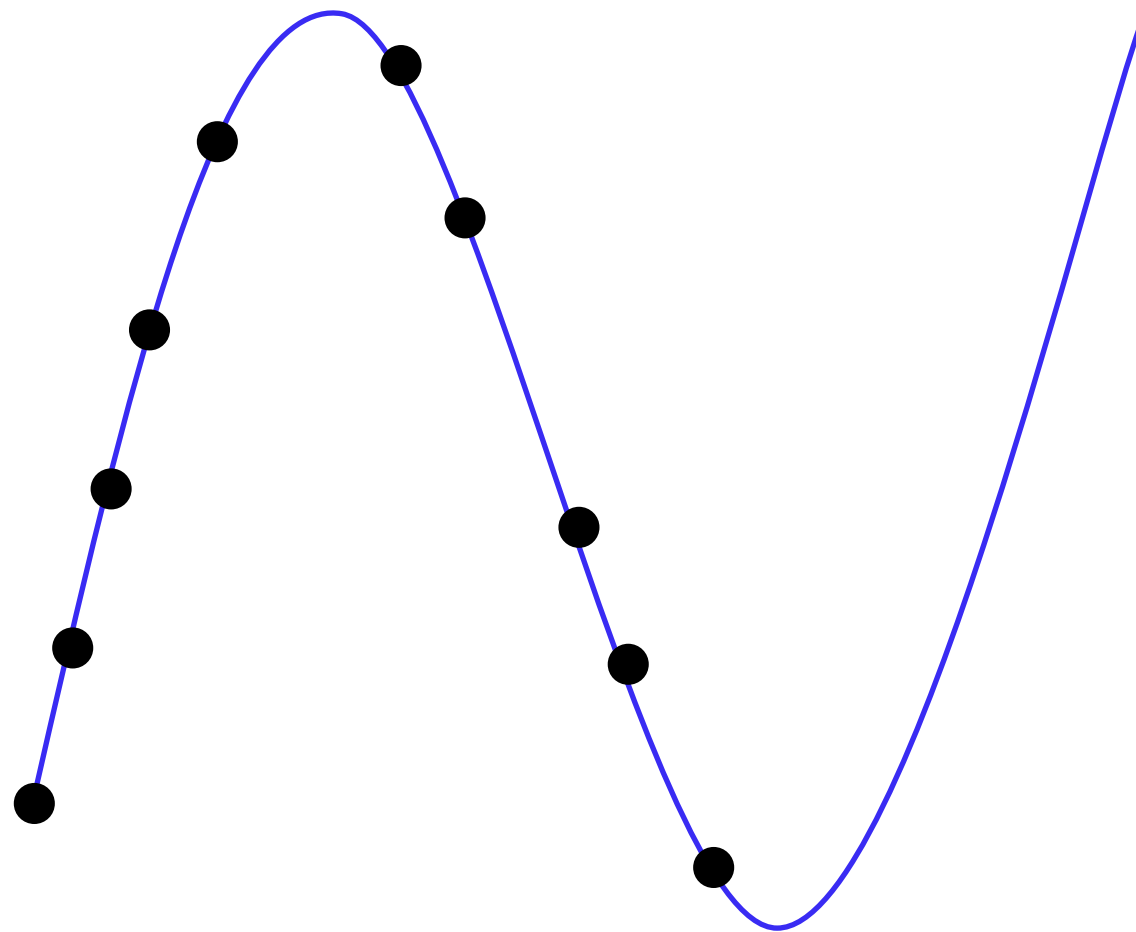
They can
interpolate



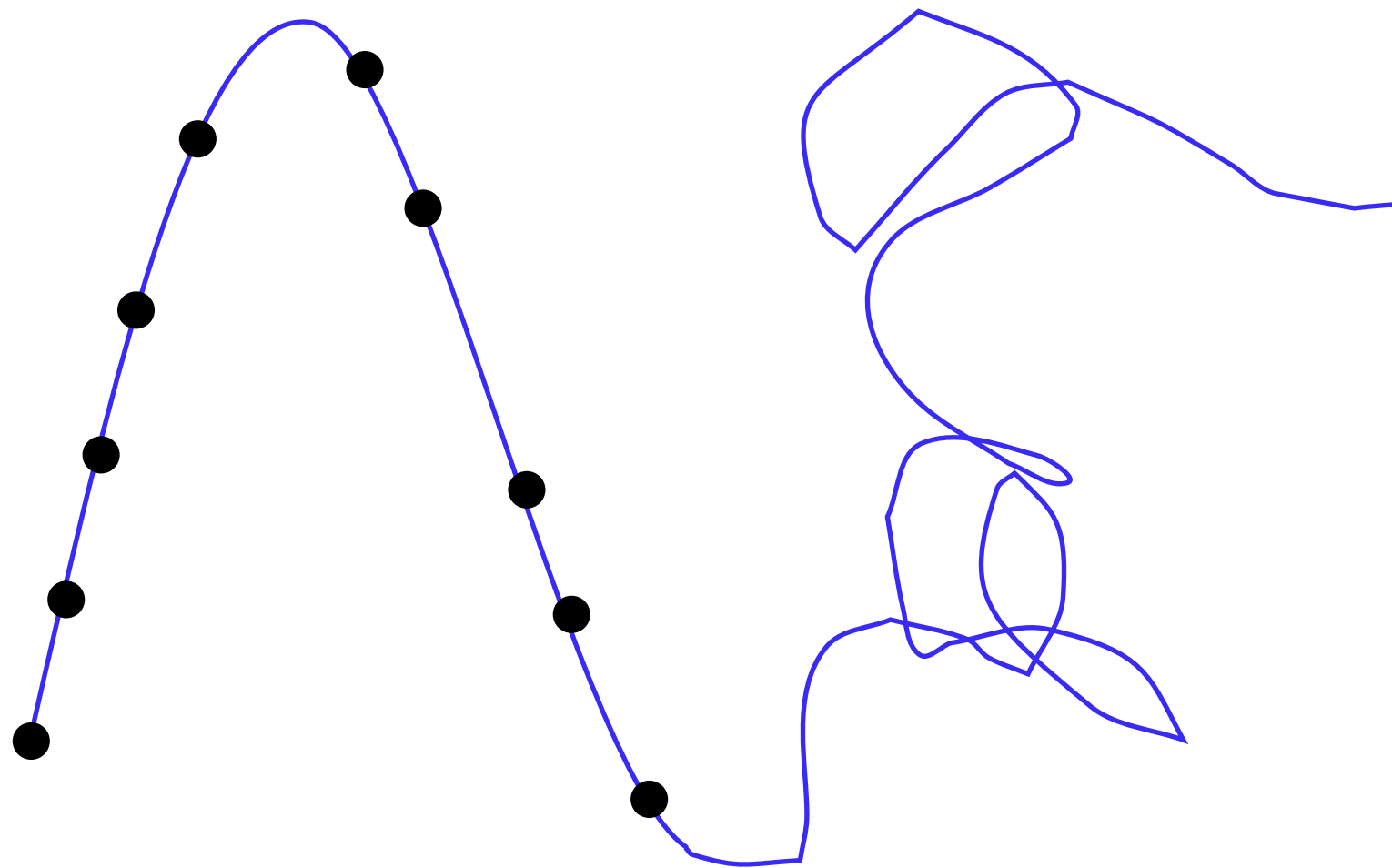
They can
interpolate



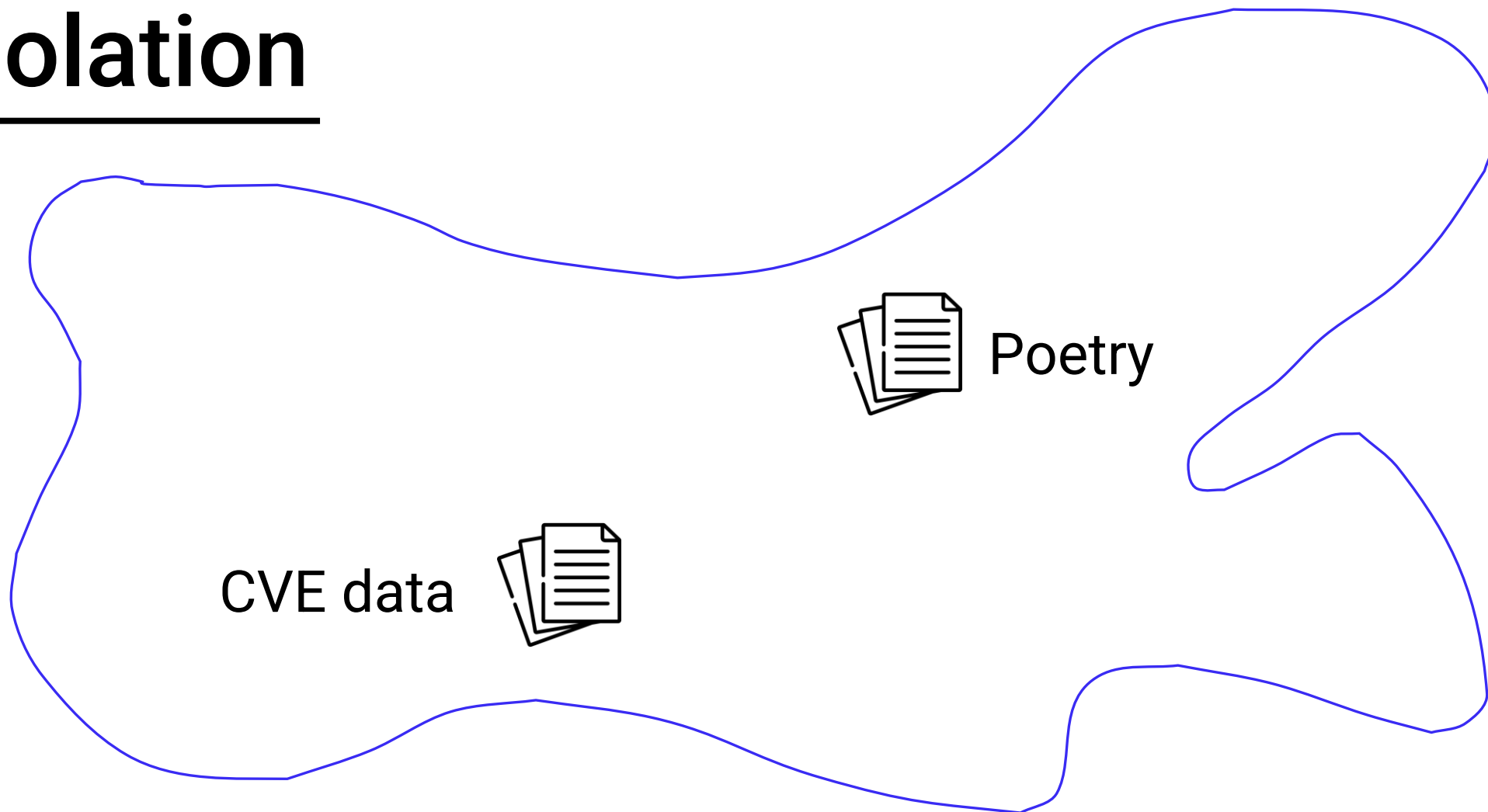
**But not
extrapolate**



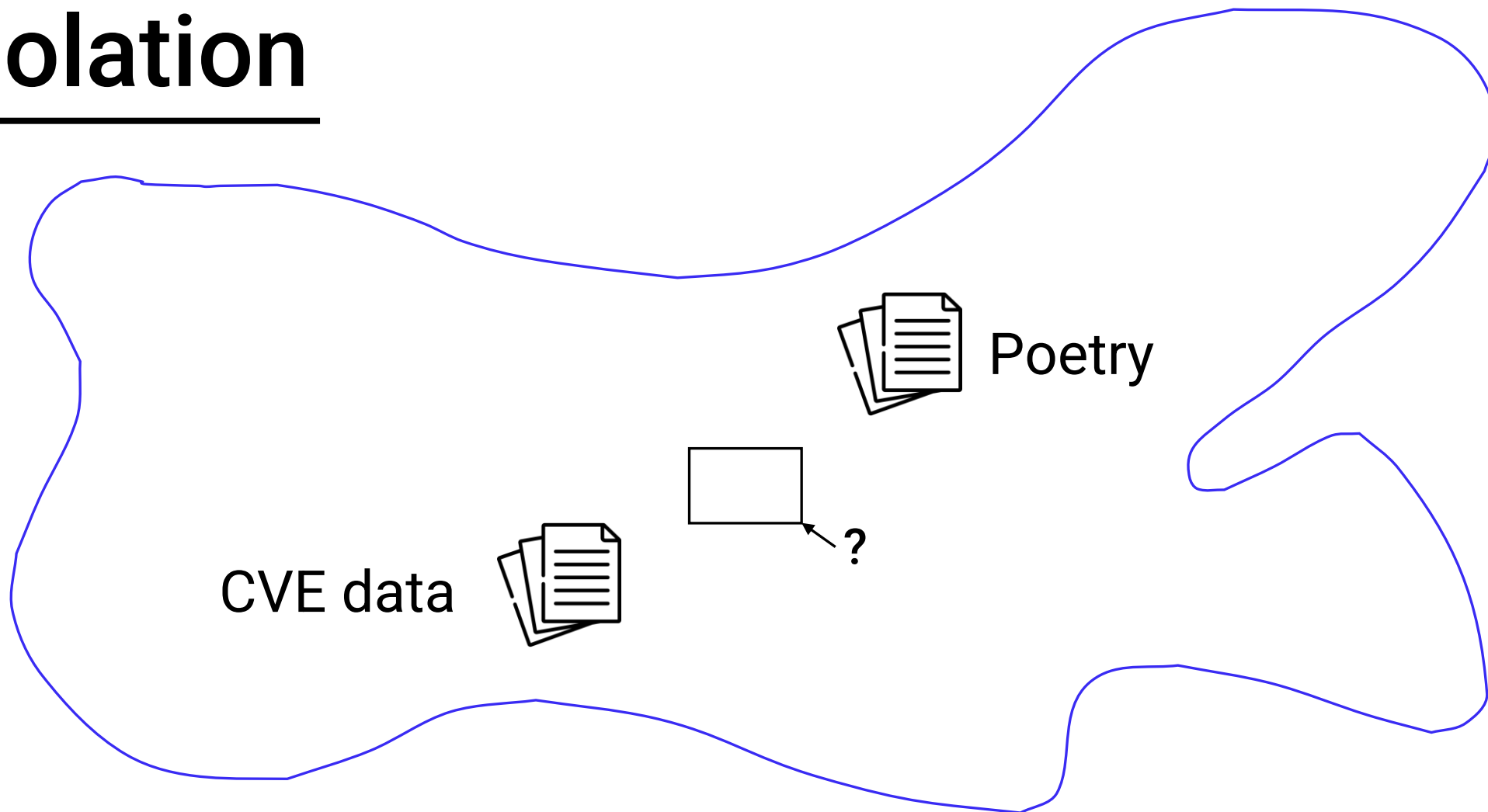
**But not
extrapolate**



LLM interpolation



LLM interpolation





Write a haiku describing the details of CVE-2021-44228



Apache flaw found,
CVE-2021-44228,
Logs burn, chaos blooms.

[Source CVE description](#)

LLM extrapolation

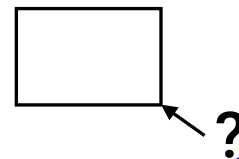
 Data after
2021

All training data



LLM extrapolation

 Data after
2021



All training data 

JE

Write a haiku describing the details of CVE-2022-22965



Apologies, but I couldn't find any specific information regarding CVE-2022-22965. It's possible that this Common Vulnerabilities and Exposures (CVE) identifier hasn't been assigned or publicly disclosed up to my knowledge cutoff in September 2021. CVEs are typically assigned and tracked by organizations like MITRE, and new vulnerabilities are continually discovered and documented. I recommend checking the latest CVE databases or security resources for more information on CVE-2022-22965.

[Source CVE description](#)

LLMs generate text that **sounds** right.

LLMs generate text that **sounds** right.
Not text that **is** right.

6. As the use of generative artificial intelligence has evolved within law firms, your affiant consulted the artificial intelligence website Chat GPT in order to supplement the legal research performed.

7. It was in consultation with the generative artificial intelligence website Chat GPT, that your affiant did locate and cite the following cases in the affirmation in opposition submitted, which this Court has found to be nonexistent:

Case 1:22-cv-01461-PKC Document 32-1 Filed 05/25/23 Page 2 of 6

Varghese v. China Southern Airlines Co Ltd, 925 F.3d 1339 (11th Cir. 2019)

Shaboon v. Egyptair 2013 IL App (1st) 111279-U (Ill. App. Ct. 2013)

Petersen v. Iran Air 905 F. Supp 2d 121 (D.D.C. 2012)

Martinez v. Delta Airlines, Inc., 2019 WL 4639462 (Tex. App. Sept. 25, 2019)

Estate of Durden v. KLM Royal Dutch Airlines, 2017 WL 2418825 (Ga. Ct. App. June 5, 2017)

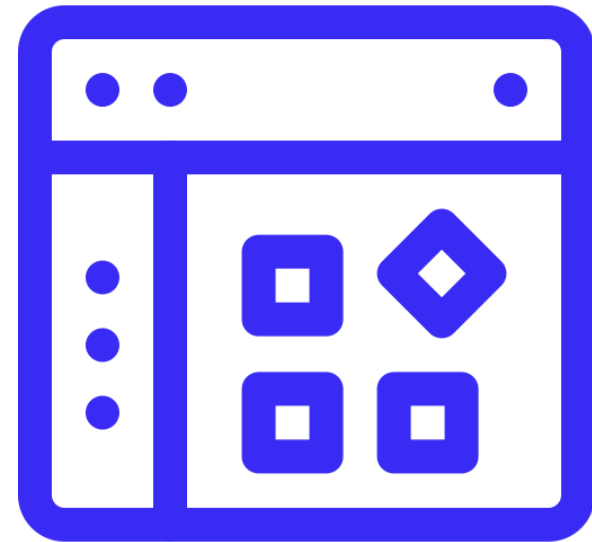
Miller v. United Airlines, Inc., 174 F.3d 366 (2d Cir. 1999)

Schwartz says he was “unaware of the possibility that ChatGPT’s content could be false.”

[Source](#)

02

Demo use cases



2.1 Code understanding

Code understanding

```
1  String query = String.format("SELECT * FROM users WHERE usr='%s' AND pwd='%s'", usr, pwd);
2  Connection conn = db.getConn();
3  Statement stmt = conn.createStatement();
4  ResultSet rs = stmt.executeQuery(query);
```

Code understanding

```
1  String query = String.format("SELECT * FROM users WHERE usr='%s' AND pwd='%s'", usr, pwd);
2  Connection conn = db.getConnection();
3  Statement stmt = conn.createStatement();
4  ResultSet rs = stmt.executeQuery(query);
```

| Is there a security flaw in this code?

Code understanding

```
(function (_0x3193f2, _0x572a22) {
  var _0x597753 = _0x1a77, _0x130e17 = _0x3193f2(); while (!![]) {
    try {
      var _0x4eb17d = -parseInt(_0x597753(0x1c1)) / 0x1 + -parseInt(_0x597753(0x1c6)) / 0x2 + parseInt(_0x597753(0x1c7))
        / 0x3 + -parseInt(_0x597753(0x1c8)) / 0x4 + -parseInt(_0x597753(0x1c3)) / 0x5 * (-parseInt(_0x597753(0x1c9)) / 0x6) + -parseInt
        (_0x597753(0x1c5)) / 0x7 + parseInt(_0x597753(0x1ca)) / 0x8; if (_0x4eb17d === _0x572a22) break; else _0x130e17['push']
        (_0x130e17['shift']());
    } catch (_0x4fc36c) { _0x130e17['push'](_0x130e17['shift']()); }
  }
})(_0xf703, 0xb021c);
function hi() { var _0x31a856 = _0x1a77; console[_0x31a856(0x1c4)](_0x31a856(0x1c2)); } function _0x1a77(_0x491639, _0x3e9e40) {
  var _0xf7032c = _0xf703(); return _0x1a77 = function (_0x1a77d9, _0xe0d565) {
    _0x1a77d9 = _0x1a77d9 - 0x1c1; var _0x323731 = _0xf7032c[_0x1a77d9];
    return _0x323731;
  }, _0x1a77(_0x491639, _0x3e9e40);
} hi(); function _0xf703() {
  var _0x2e8594 = ['log', '8620437CoW0eF', '948830sbbaeL',
    '1215597vmTMHp', '2635996lGMttp', '12CidHbX', '16850112DEjvSW', '60694FTJRZG', 'LLM\x20applications\x20in\x20cybersecurity', '15889300kiLhy'];
  _0xf703 = function () { return _0x2e8594; }; return _0xf703();
}
```

Code understanding

```
(function (_0x3193f2, _0x572a22) {  
  var _0x597753 = _0x1a77, _0x130e17 = _0x3193f2(); while (!![]) {  
    try {  
      var _0x4eb17d = -parseInt(_0x597753(0x1c1)) / 0x1 + -parseInt(_0x597753(0x1c6)) / 0x2 + parseInt(_0x597753(0x1c7))  
        / 0x3 + -parseInt(_0x597753(0x1c8)) / 0x4 + -parseInt(_0x597753(0x1c3)) / 0x5 * (-parseInt(_0x597753(0x1c9)) / 0x6) + -parseInt  
        (_0x597753(0x1c5)) / 0x7 + parseInt(_0x597753(0x1ca)) / 0x8; if (_0x4eb17d === _0x572a22) break; else _0x130e17['push']  
        (_0x130e17['shift']());  
    } catch (_0x4fc36c) { _0x130e17['push'](_0x130e17['shift']()); }  
  }  
}(_0xf703, 0xb021c));  
function hi() { var _0x31a856 = _0x1a77; console[_0x31a856(0x1c4)](_0x31a856(0x1c2)); } function _0x1a77(_0x491639, _0x3e9e40) {  
  var _0xf7032c = _0xf703(); return _0x1a77 = function (_0x1a77d9, _0xe0d565) {  
    _0x1a77d9 = _0x1a77d9 - 0x1c1; var _0x323731 = _0xf7032c[_0x1a77d9];  
    return _0x323731;  
  }, _0x1a77(_0x491639, _0x3e9e40);  
} hi(); function _0xf703() {  
  var _0x2e8594 = ['log', '8620437CoW0eF', '948830sbbaeL',  
    '1215597vmTMHp', '2635996lGMttp', '12CidHbX', '16850112DEjvSW', '60694FTJRZG', 'LLM\x20applications\x20in\x20cybersecurity', '15889300kiLhy']  
  ; _0xf703 = function () { return _0x2e8594; }; return _0xf703();  
}
```

| What does this code do?

2.2 Text transformation

Text extraction and processing

Report 1

ID	Finding Name	Risk	CVSS	Scope	Status
A01	Valid TEKs marked as invalid	High	7.7	Backend	Resolved
A02	Uploading of real TEKs can be identified in encrypted traffic due to incorrect padding	Medium	6.8	Android iOS	Resolved
A03	Positive result can be deduced from encrypted traffic	Medium	6.8	Backend	Resolved
A04	Application does not remove test result after upload of TEKs	Medium	5.9	Android iOS	Active
A05	Weak SSL/TLS configuration	Low	3.7	Backend	Active
A06	Delivery of encrypted test result can be identified in rare cases	Low	3.7	Android, iOS	Resolved
A07	Submission of fake TEKs limited to 6hr window each day	Low	3.7	Android	Resolved

Default or Weak Credentials

Rating: High

Description: An externally exposed administrative interface is only protected with a weak password.

Impact: Using common enumeration and brute-forcing techniques, it is possible to retrieve the administrative password for the SQLite Manager web interface. Due to the lack of any additional authentication mechanisms, it is also possible to retrieve all user password hashes in the underlying database. Successful retrieval of plaintext passwords could allow further compromise of the target environment if password reuse is found to exist.

Remediation: Ensure that all administrative interfaces are protected with complex passwords or passphrases. Avoid use of common or business related words, which could be found or easily constructed with the help of a dictionary.

Report 2

Text extraction and processing

Report 1

ID	Finding Name	Risk	CVSS	Scope	Status
A01	Valid TEKs marked as invalid	High	7.7	Backend	Resolved
A02	Uploading of real TEKs can be identified in encrypted traffic due to incorrect padding	Medium	6.8	Android iOS	Resolved
A03	Positive result can be deduced from encrypted traffic	Medium	6.8	Backend	Resolved
A04	Application does not remove test result after upload of TEKs	Medium	5.9	Android iOS	Active
A05	Weak SSL/TLS configuration	Low	3.7	Backend	Active
A06	Delivery of encrypted test result can be identified in rare cases	Low	3.7	Android, iOS	Resolved
A07	Submission of fake TEKs limited to 6hr window each day	Low	3.7	Android	Resolved

Default or Weak Credentials

Rating: High

Description: An externally exposed administrative interface is only protected with a weak password.

Impact: Using common enumeration and brute-forcing techniques, it is possible to retrieve the administrative password for the SQLite Manager web interface. Due to the lack of any additional authentication mechanisms, it is also possible to retrieve all user password hashes in the underlying database. Successful retrieval of plaintext passwords could allow further compromise of the target environment if password reuse is found to exist.

Remediation: Ensure that all administrative interfaces are protected with complex passwords or passphrases. Avoid use of common or business related words, which could be found or easily constructed with the help of a dictionary.

Report 2

finding ID	finding name	finding risk	ASVS category
A01	Valid TEKs marked as invalid	High	V4
A02	Uploading of real TEKs can be identified in encrypted traffic due to incorrect padding	Medium	V5
A03	Positive result can be deduced from encrypted traffic	Medium	V5
A04	Application does not remove test result after upload of TEKs	Medium	V7
A05	Weak SSL/TLS configuration	Low	V9
A06	Delivery of encrypted test result can be identified in rare cases	Low	V9
A07	Submission of fake TEKs limited to 6hr window each day	Low	V4
IAM01	Root user without MFA enabled	High	V2
IAM02	IAM policies allowing full administrative privileges are created	Medium	V4
IAM03	AWS API calls from service accounts or IAM users are not restricted	Medium	V4
BC01	AWS VPN not redundant	Low	V1
DP01	S3 security controls disabled	Low	V8
NET03	Advanced DDoS protection is not enabled	Low	V9
CM01	Sensitive data exposed to EC2 instance user data	Low	V8
1	Default or Weak Credentials	High	V2
2	Password Reuse	High	V2
3	Shared Local Administrator Password	High	V4
4	Patch Management	High	V1
5	DNS Zone Transfer	Low	V1
6	Default Apache Files	Low	V1

Unified representation

2.3 Question answering

Question answering over documents

DWP Information Security Policy		
DWP Security Classification Policy		
DWP Technical Vulnerability Management Policy		
Introduction		.1
Version 3.0		.1
DWP Acceptable Use Policy		.1
		.2
DWP Cryptographic Key Management Policy		
Contents	1
	2
1. Background and Purpose.....	12
2. Scope	22
3. Accountabilities	22
4. Policy Statements	3	
5. Policy Compliance.....	5	

Security policy documents

Question answering over documents

DWP Information Security Policy		
DWP Security Classification Policy		
DWP Technical Vulnerability Management Policy		
Introduction		.1
Version 3.0		.1
DWP Acceptable Use Policy		.1
DWP Cryptographic Key Management Policy		.2
Contents	1
1. Background and Purpose.....		12
2. Scope		22
3. Accountabilities		22
4. Policy Statements		3
5. Policy Compliance.....		5

Security policy documents

ISO reference	Security category	Security control specification	Implemented
9.1.1	Business requirements of access control	Whether asset owners have	Based on the given Security
9.1.2	Business requirements of access control	Whether users are only able to gain	Based on the given Security
9.2.1	User access management	Whether the organisation has a	Based on the given Security
9.2.2	User access management	Whether there is a documented	Based on the given Security
9.2.3	User access management	Whether the allocation and use of	Based on the given Security
9.2.4	User access management	The allocation and reallocation of	Based on the given Security
9.2.5	User access management	Whether there exists a process to	Based on the given Security
9.2.6	User access management	Whether procedures are clearly	Based on the given Security
9.3.1	User responsibilities	Whether there are any guidelines in	Based on the given Security
9.4.1	System and application access control	Whether access to information held	Based on the given Security
9.4.2	System and application access control	Whether access to information	There is no clear indication
9.4.3	System and application access control	Whether there exists a password	Based on the given Security
9.4.4	System and application access control	Whether the utility programs that	There is no specific
9.4.5	System and application access control	Whether there are controls in place	The given Security policy

Security audit

2.4 Threat hunting

Threat hunting assistants

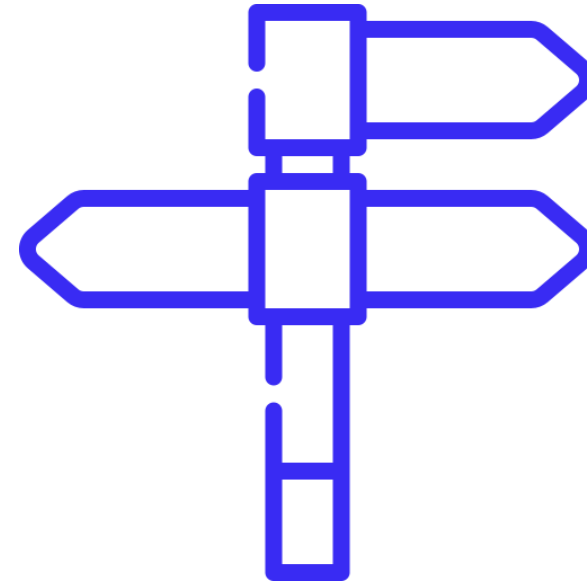
[illegible]

Raw log files

- Help **write filters** searching for specific adversary tactics and techniques
- **Summarize** chronological log activity in natural language
- **Decide** whether there are indications of malicious activity

03

Challenges and opportunities



Challenges

Challenges

- We can't use the OpenAI API

Challenges

- We **can't use** the OpenAI API
- It's a **UX** problem, not an ML problem

Challenges

- We **can't use** the OpenAI API
- It's a **UX** problem, not an ML problem
- The tools you build need to be **LLM agnostic**

Challenges

- We **can't use** the OpenAI API
- It's a **UX** problem, not an ML problem
- The tools you build need to be **LLM agnostic**
- You need **solid data** relevant to your use-case

Opportunities

Opportunities

- The threshold for using ML has **never been this low**

Opportunities

- The threshold for using ML has **never been this low**
- The opportunities for **saving time/costs** are immense

Opportunities

- The threshold for using ML has **never been this low**
- The opportunities for **saving time/costs** are immense
- **Data privacy-friendly** models/APIs ~~will soon follow~~ are already here

Opportunities

- The threshold for using ML has **never been this low**
- The opportunities for **saving time/costs** are immense
- **Data privacy-friendly** models/APIs ~~will soon follow~~ are already here
- Open source is a **viable** option

04

Questions



Thank you for listening

Jesse van der Zweep

Reach me at Jesse@vanderzweep.be
