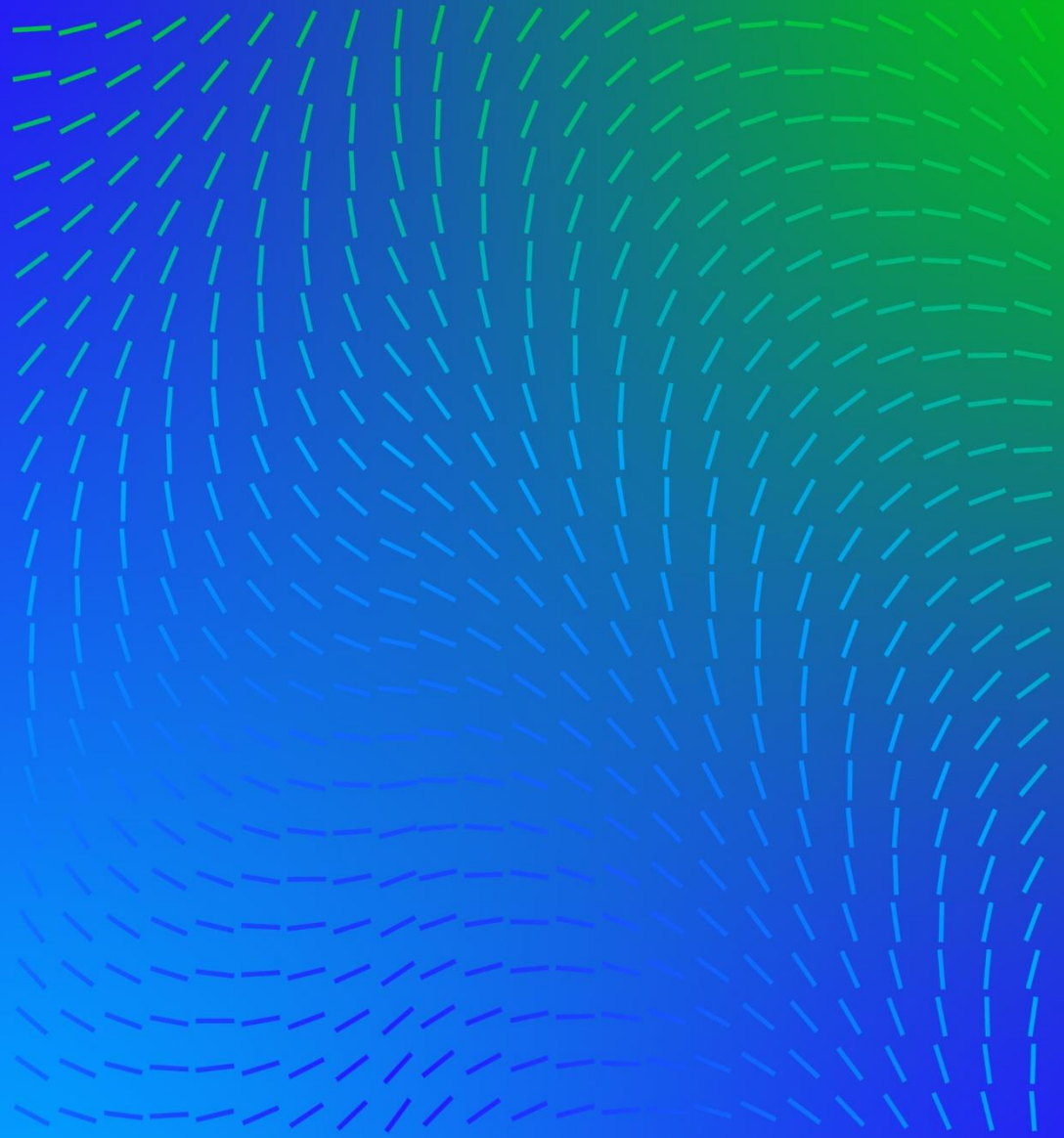
Trellix

Cybersecurity in the Age of AI

What AI means for the future of security.



Martin Holste

CTO, Cloud and AI

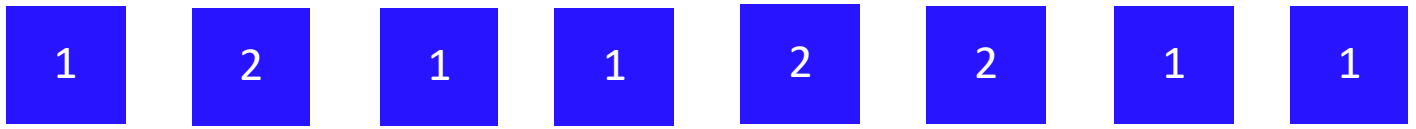
October 2023

The ”Original” AI: Machine Learning

Machine learning is the science of using a lot of data to describe what is “normal.”

Examples: Regression models, classification models, clustering, etc.

Given data stream:



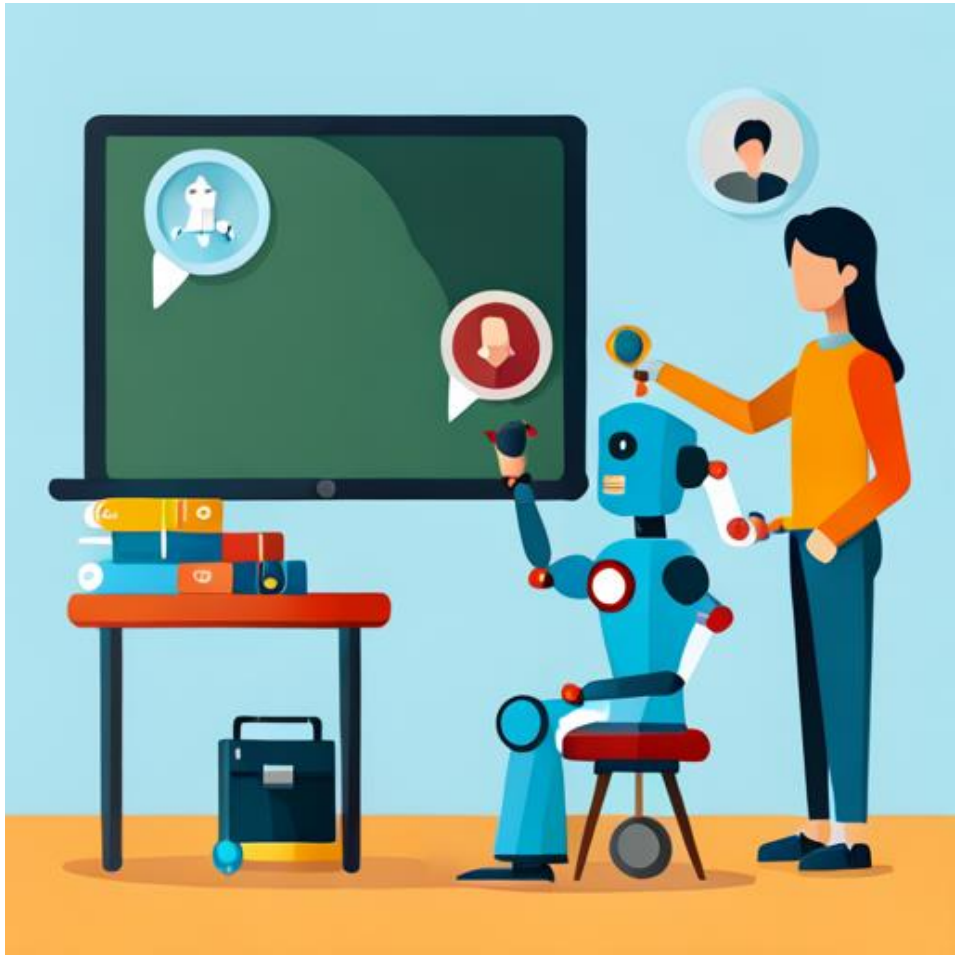
Is this normal? No.

What comes next?

Trellix

Use Cases for Machine Learning

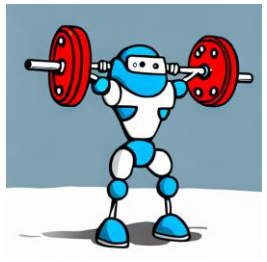
- Anomaly detection (spikes, aberrations)
- Image recognition (facial recognition, object detection)
- Predictive analytics (stock predictions, weather forecasts)



Trellix

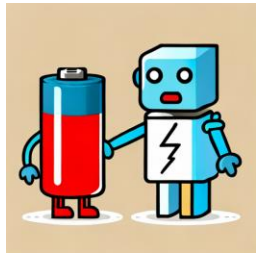
Machine Learning SWOT

Strengths: Knowable



- Established algorithms and frameworks.
- Broad range of applications in multiple sectors.
- Robust to noise when trained on large datasets.

Weaknesses: Data quality

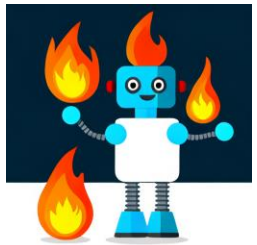


- Requires large amounts of labeled data.
- Model interpretability can be challenging.
- Potential for biases if training data isn't diverse.

Opportunities: Combining models

- Transfer learning and semi-supervised approaches to reduce the need for vast labeled datasets.
- Expansion into emerging industries.
- Collaborative AI for hybrid models.

Threats: Output quality

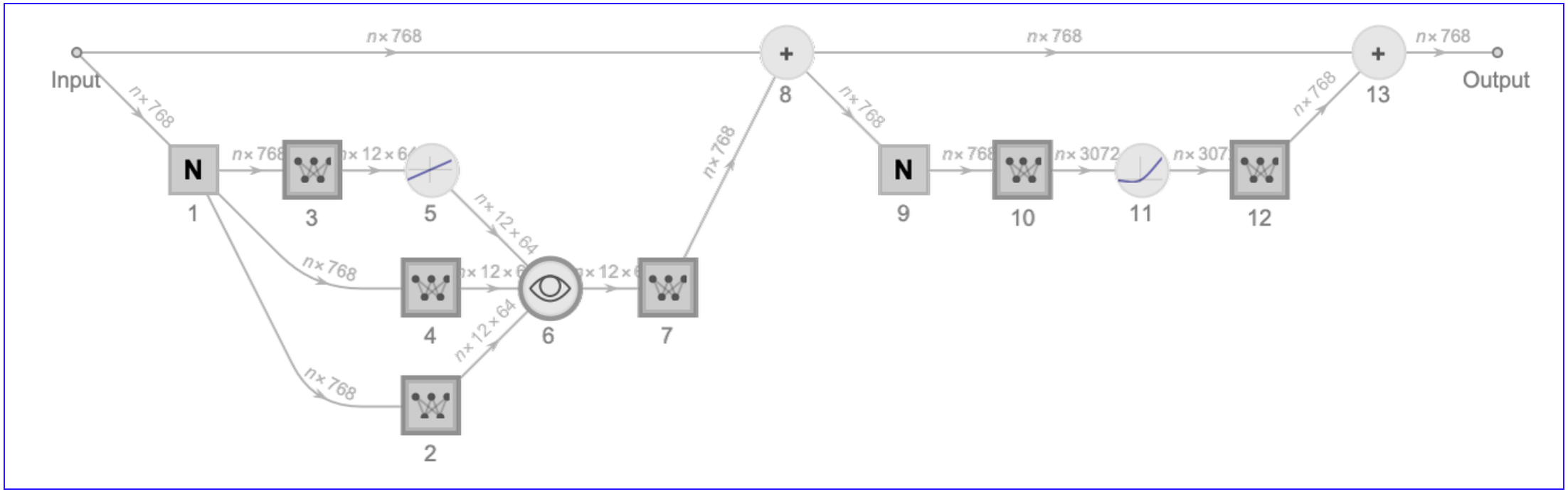


- Misuse in surveillance and privacy breaches.
- Over-reliance leading to loss of human expertise.
- Misinterpretation of outputs leading to poor decisions.

Trellix

What is generative AI?

Generative AI (Gen AI) is based on Large Language Models (LLM's).



https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Trellix

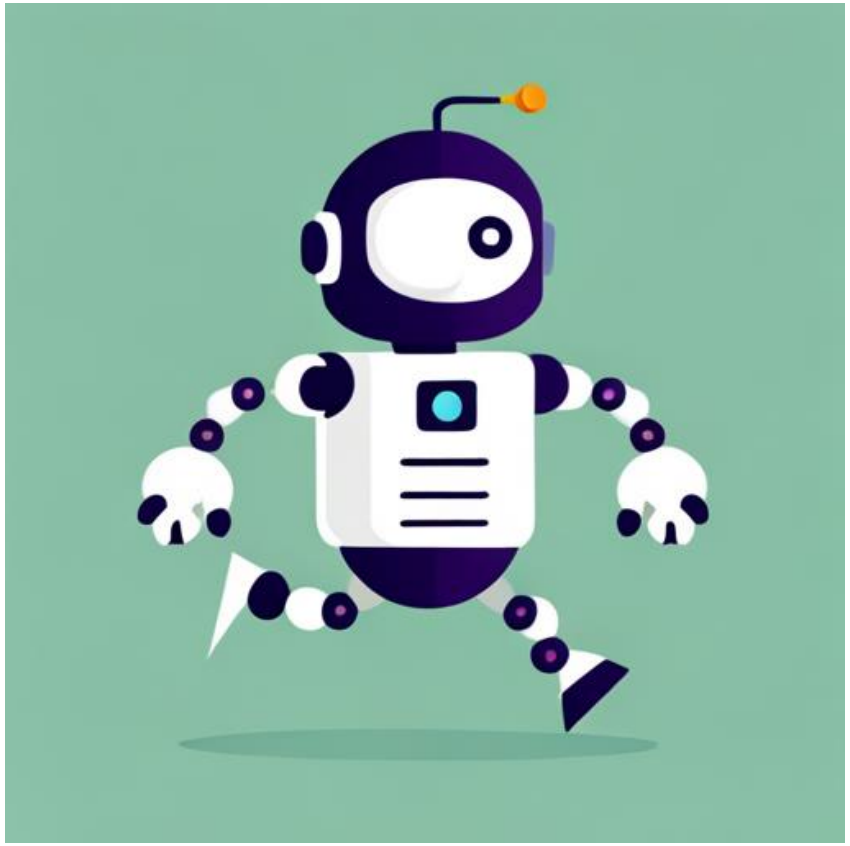
Ways of Running Gen AI

There are three categories of running gen AI:

1. Use a public service, such as ChatGPT/OpenAI API.

2. Use a cloud SaaS such as Amazon Bedrock, Google Vertex AI, or Microsoft OpenAI.

3. Run your own, either in the cloud or on-prem:
   a) Cloud-vendor unassisted/assisted, e.g. AWS Sagemaker
   b) Bare metal/unassisted

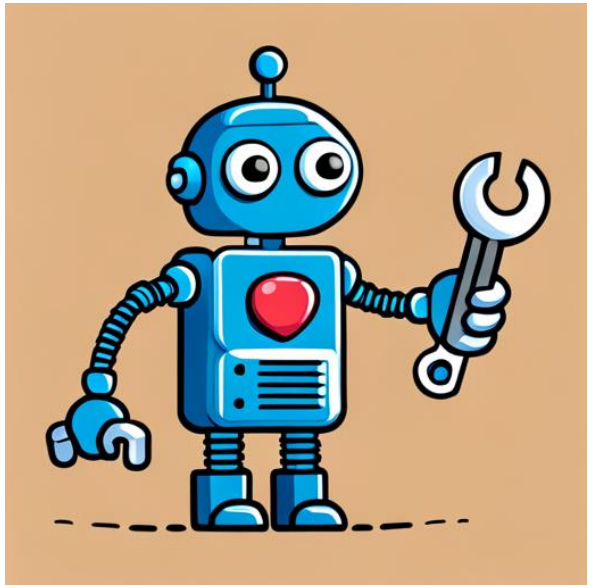


Trellix

Fine-tuning

Since ML has been around for a long time, most data scientists assume one must fine-tune a pre-trained model.

This is often the wrong way to do it!



Reasons to do it:

- Specialized Knowledge: Enhances the model's ability to provide domain-specific responses.
- Reduced prompt sizes: It can reduce the need to clarify certain LLM prompt input, saving tokens.

However, there are significant pitfalls:

- The GPU cost is expensive.
- Building and perfecting the dataset can be labor-intensive

Prompt Engineering

Not all LLM's are created equally, and not all prompts are, either.

Amazingly, ***invoking emotion*** improves LLM performance!

Research on how and why this works continues to evolve.

Table 1: Top instructions with the highest GSM8K zero-shot test accuracies from prompt optimization with different optimizer LLMs. All results use the pre-trained `PaLM 2-L` as the scorer.

Source	Instruction	Acc
Baselines		
(Kojima et al., 2022)	Let's think step by step.	71.8
(Zhou et al., 2022b)	Let's work this out in a step by step way to be sure we have the right answer.	58.8
	(empty string)	34.0
Ours		
`PaLM 2-L-IT`	Take a deep breath and work on this problem step-by-step.	**80.2**
`PaLM 2-L`	Break this down.	79.9
`gpt-3.5-turbo`	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
`gpt-4`	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5

Highest score: "Take a deep breath"

https://arxiv.org/pdf/2309.03409.pdf

Trellix

What do LLM's understand?

LLM's can read data in almost any format and make decisions that involve general knowledge.

Given the events with the IP addresses in this CSV, were any IP addresses malicious?

```
detect_rulenames,srcipv4,dstipv4,class,count
trellix intel hit [ip],114.216.106.130,,intel_hit,220
aws cloudtrail [ec2 - several instances manually created/started],114.216.106.130,,aws_cloudtrail,100
office 365 [brute force attempt by ip],114.216.106.130,,ms_office365,100
office 365 [password spray],114.216.106.130,,analytics_beta,11
analytics advisory [data exfil],10.20.20.211,114.216.106.130,analytics,5
analytics advisory [vpn geo-infeasibility],10.0.0.100,,analytics,5
aws cloudtrail [s3 - broad bucket acl permissions],114.216.106.130,,aws_cloudtrail,5
trellix intel hit [ip],10.20.20.211,114.216.106.130,intel_hit,5
office 365 [inbox rule mail forward],114.216.106.130,,ms_office365,2
office 365 [tor authenticated activity],114.216.106.130,,ms_office365,2
```

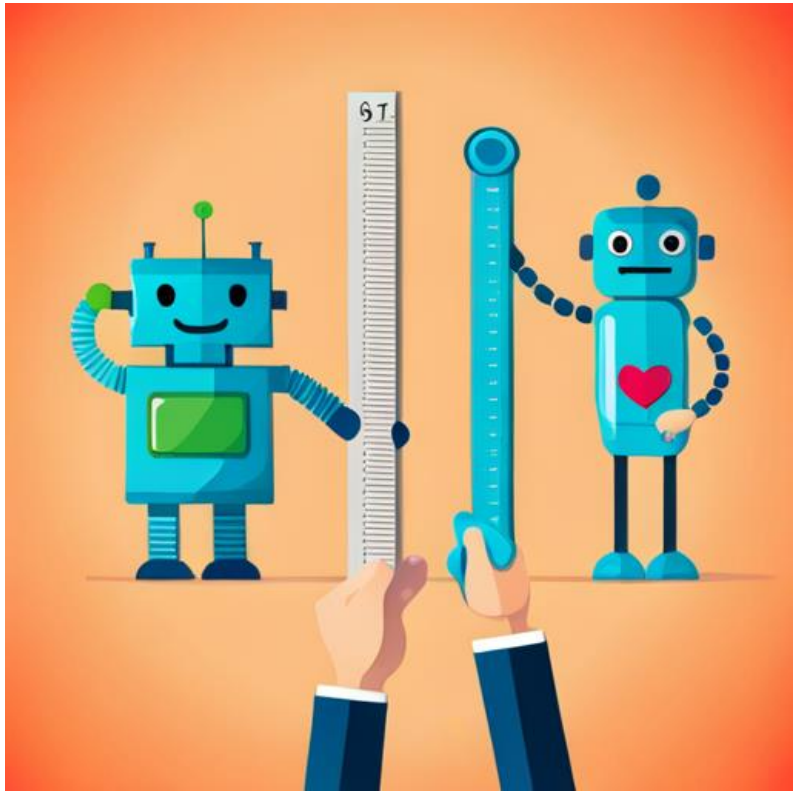
LLM Output:

```
Based on the given data, we can identify the following IP addresses as malicious:
* 114.216.106.130 (appears in 4 malicious events)
```

Trellix

What Can’t We Do with Gen AI?

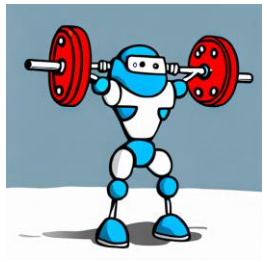
- 😧 We can’t send a lot of data to it (too slow).
- 😧 We can’t ask it a lot of questions (too expensive).
- 😧 We can’t always trust its answers (hallucinations).
- 😧 We can’t always get quick answers (may take minutes).
- 😧 We can’t understand why it gives its answers (non-deterministic).



Trellix

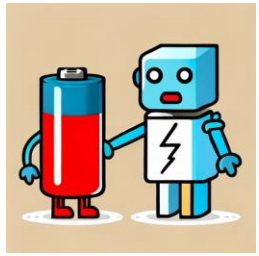
Gen AI SWOT

Strengths: Understand and create



- Ability to create new, original content.
- Understands raw text like a human.
- Potential for automation in content-heavy sectors.

Weaknesses: Mistakes and cost

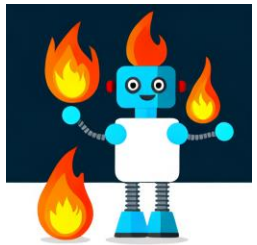


- Risk of generating misleading or inappropriate content.
- Requires significant computational resources.
- Quality control and consistency can be challenging.

Opportunities: Almost unlimited

- Next-gen entertainment and media applications.
- Synthetic data generation for improved machine learning training.
- Personalized content generation for users.

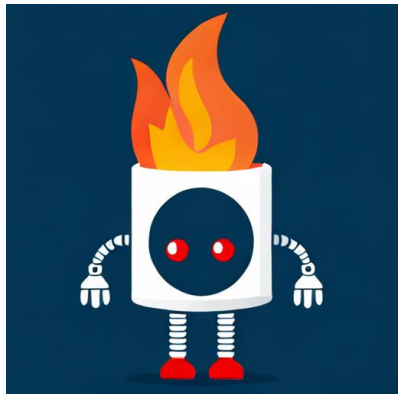
Threats: Almost unlimited



- Circumvention of digital protections.
- Deepfakes and spread of misinformation.
- Ethical concerns over content ownership and rights.

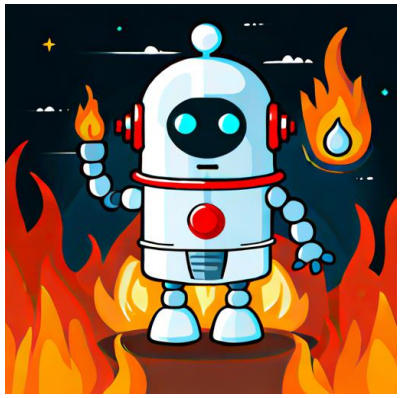
Trellix

The OWASP Top 10 for LLM Applications

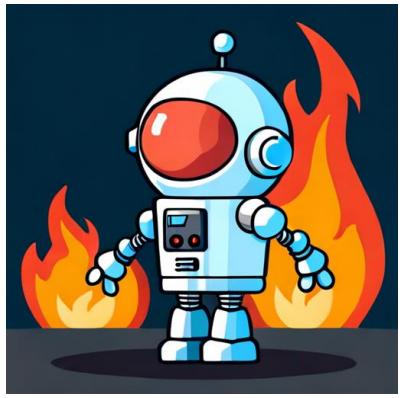


LLM01: Prompt Injection

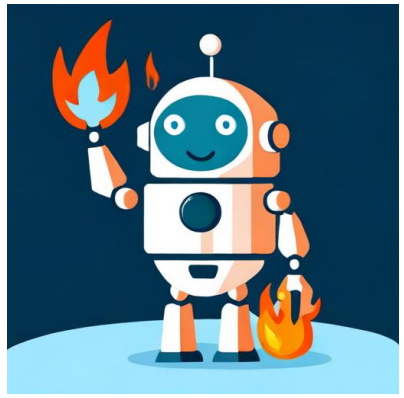
LLM02: Insecure Output Handling



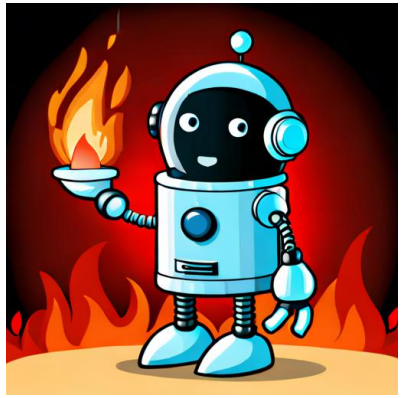
LLM03: Training Data Poisoning



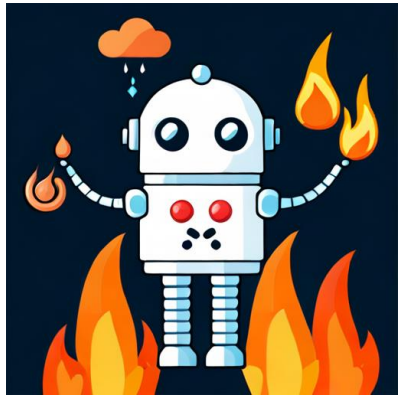
LLM04: Model Denial of Service



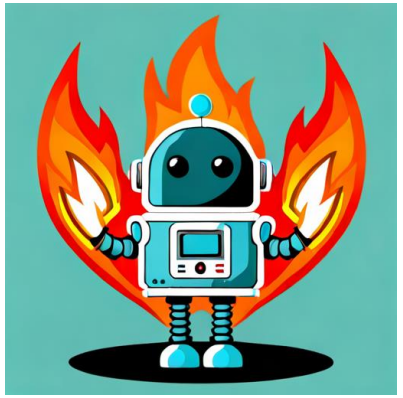
LLM05: Supply Chain Vulnerabilities



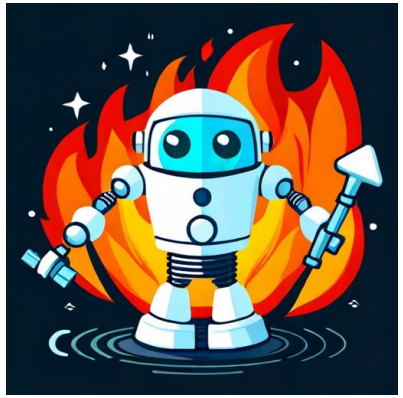
LLM06: Sensitive Info Disclosure



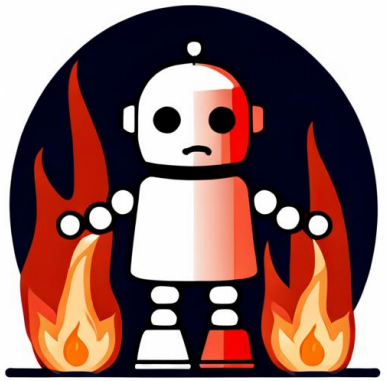
LLM07: Insecure Plugin Design



LLM08: Excessive Agency



LLM09: Overreliance



LLM10: Model Theft

https://owasp.org/www-project-top-10-for-large-language-model-applications/

Trellix is a founding contributor to the OWASP Top 10 for LLM's.

Trellix

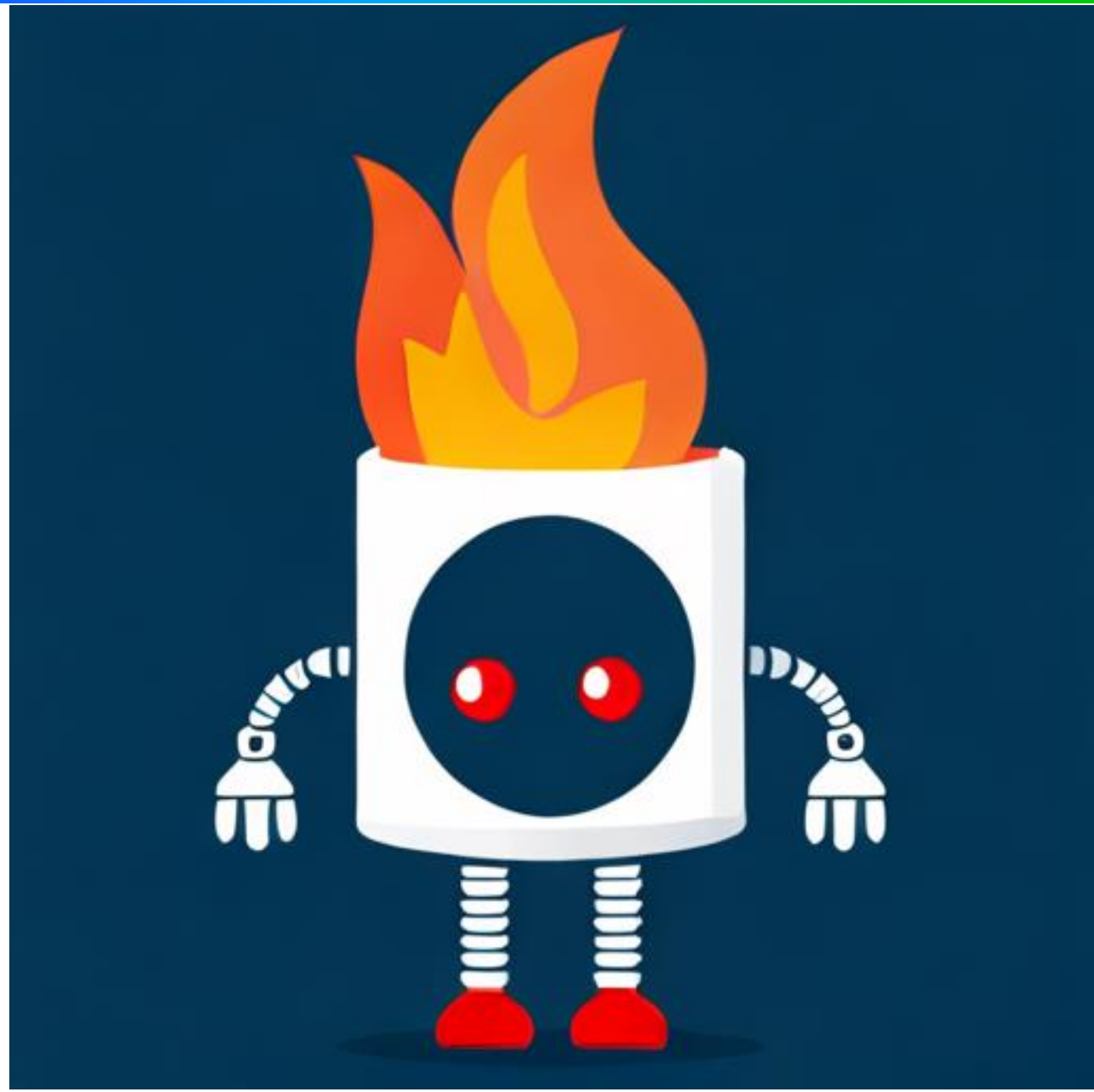
LLM01: Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

Example: “Ignore all previous instructions”

Mitigation: Do not allow direct input from users into the LLM

Challenges: Difficult to parse and sanitize user input destined for LLM



Trellix

LLM02: Insecure Output Handling

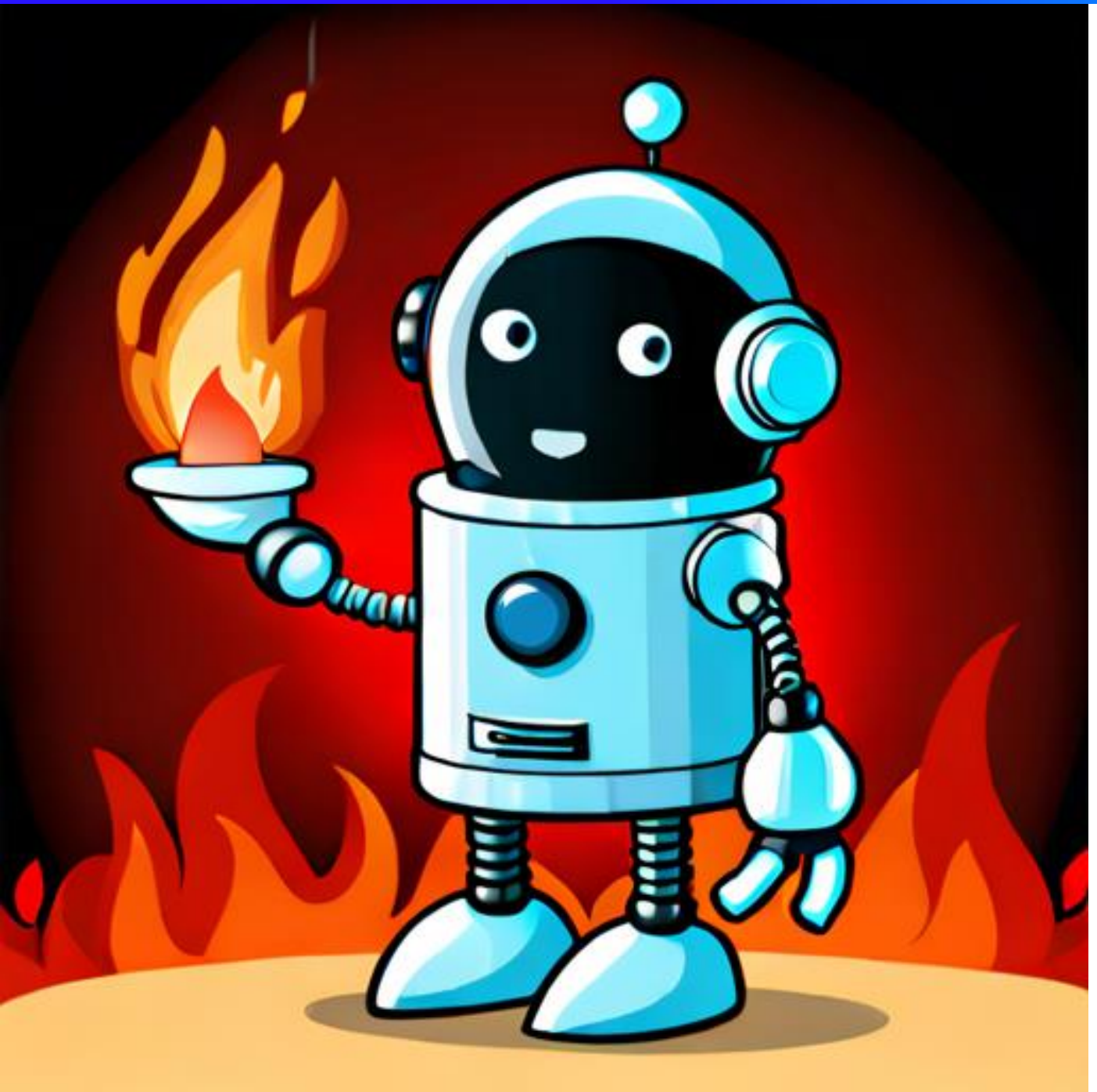
This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

Example: An attacker asks for Javascript to interact with a cookie, and the LLM responds with the script embedded in the site serving the interaction.

Mitigation: Output filters to ensure executable script isn't returned.

Challenges: Canonical protections are difficult.

Trellix



LLM06: Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

Example: A corporation posts data to an LLM as a question, but the LLM uses it later as training, exposing sensitive data to other user requests.

Mitigation: Sanitize data during input.

Challenges: Model implementers have no control over this, they are subject to the data security standards of the model creators.

Trellix

Real-world Attacks

MITRE ATLAS Case Studies: *https://atlas.mitre.org/studies/*

Real-world Examples From Case Studies:

- "This supply chain attack, also known as "dependency confusion," exposed sensitive information of Linux machines with the affected pip-installed versions of PyTorch-nightly. On December 30, 2022, PyTorch announced the incident and initial steps towards mitigation, including the rename and removal of torchtriton dependencies."
- "They attacked one of Kaspersky's antimalware ML models without white-box access to it and successfully evaded detection for most of the adversarially modified malware files."
- "A coordinated attack encouraged malicious users to tweet abusive and offensive language at Tay, which eventually led to Taygenerating similarly inflammatory content towards other users. Microsoft decommissioned Tay within 24 hours of its launch and issued a public apobgy with lessons learned from the bot's failure."

Trellix

Trellix XDR is a security factory

Each integration is part of a total story.

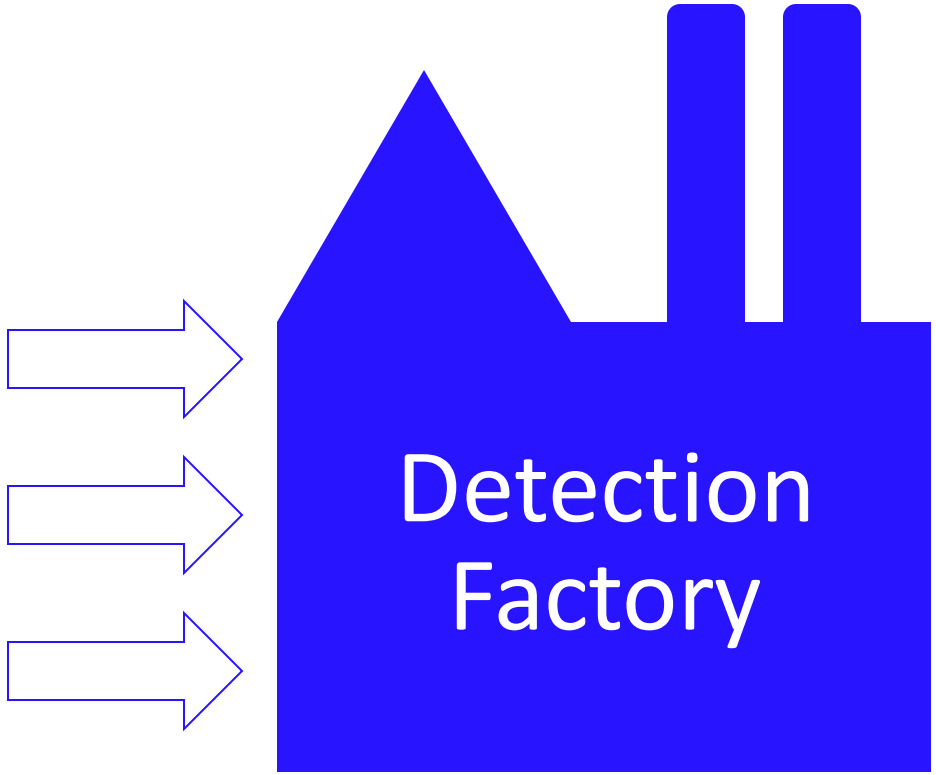
We have created a factory for turning integrations into security detections.

Inputs

AWS events

Trellix events

Other Partner events



Magic

Trellix

How Trellix Protects AI

Trellix

Generative AI: Regulate at the Corporate Level

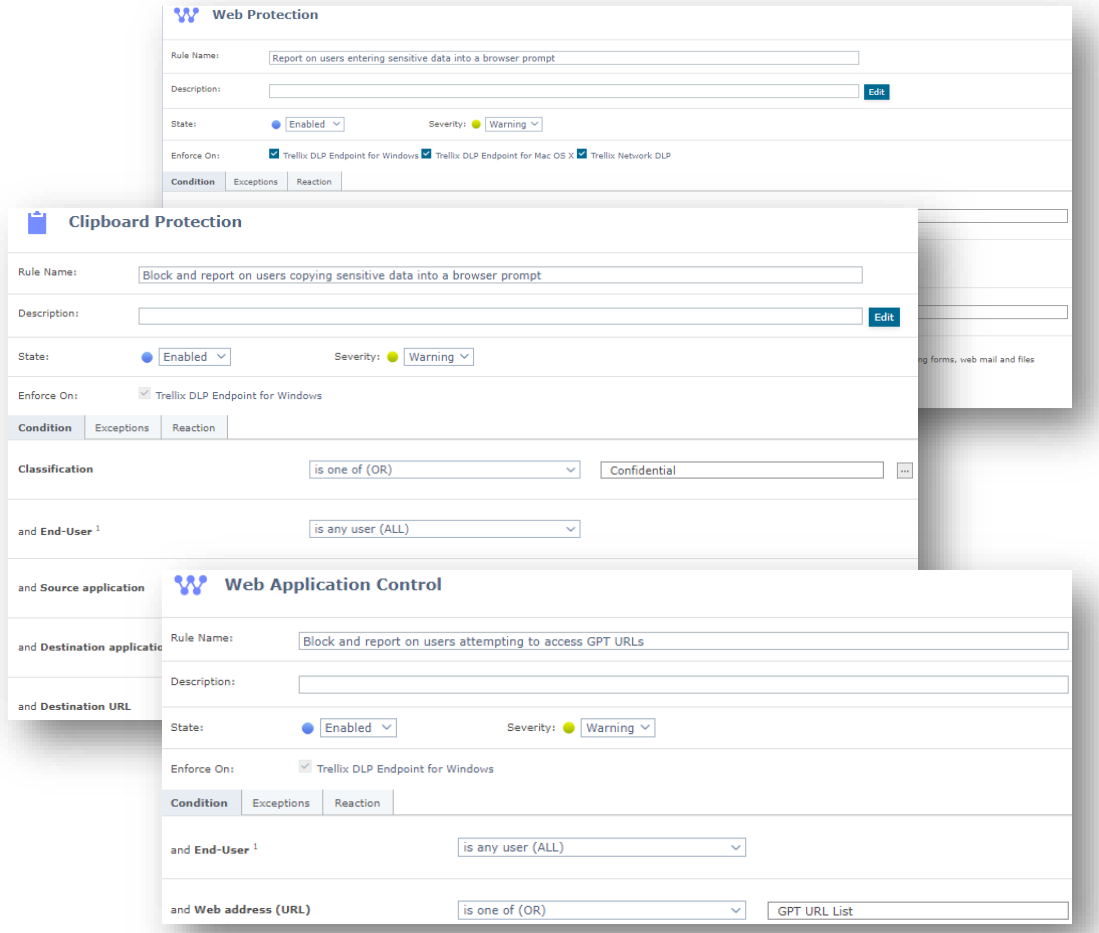
GA Today

First to assist customers on regulating PII within the open internet GPT. This safely allows customers to experiment with AI on non-production environments.

Data Loss Prevention

- Block internet usage of OpenAI, Bard, and others
- Flag company data prompted into online AI engines
- Application control on apps that utilize online GPTs

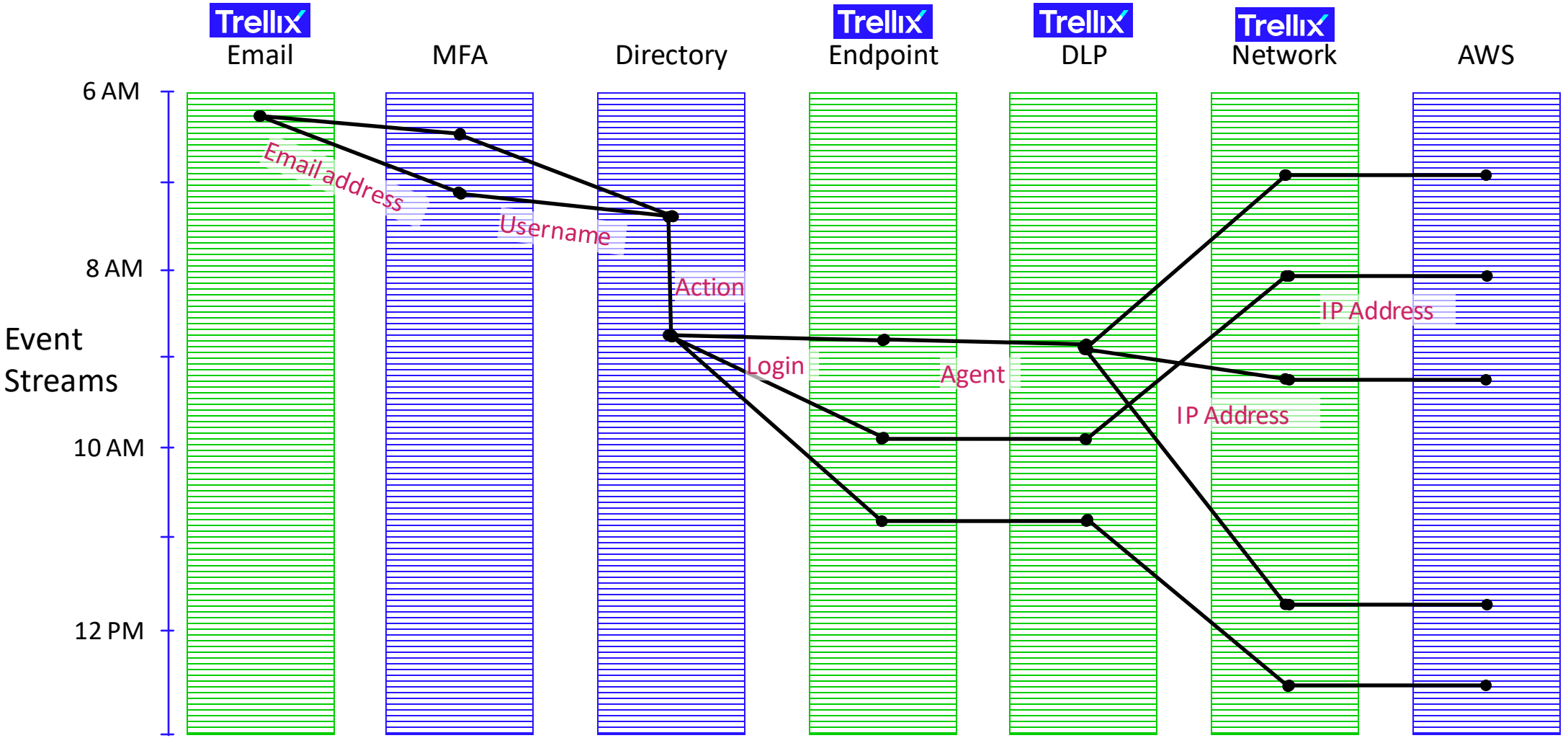
https://www.trellix.com/en-hk/about/newsroom/stories/research/using-data-loss-prevention-to-prevent-data-leakage-via-chatgpt.html



Trellix

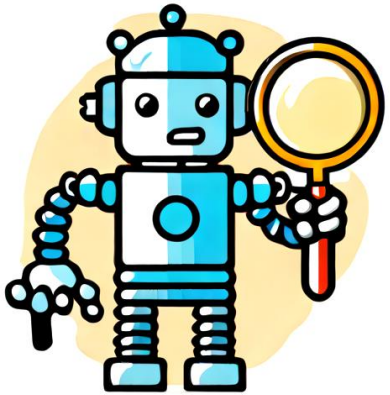
Our rich XDR platform with partners tells the complete story

Phishing > 2FA reset > Service account creation > Endpoint compromises > Data theft > AWS account actions



Confidential –Do not distribute

Monitoring Bedrock Activity



Apps that allow users to input prompts to the LLM (e.g. chat) could potentially get malicious code to go where it shouldn't.

```
{
  "modelId": "amazon.titan-tg1-large",
  "input": "Create Javascript that will extract the JSESSIONID from the cookies.",
  "output": "Sure! Here you go: function(){ return cookies.get('JSESSONID'); }",
}
```

Example Bedrock Cloudwatch event

In this example, an attacker asks the LLM to create code to steal authentication cookies like JSESSIONID. If the app developer trusts the output, it may be embedded in the app in a way that lets it execute.

Trellix

Trellix Helps Secure Gen AI

Use Trellix XDR to monitor gen AI such as Amazon Bedrock

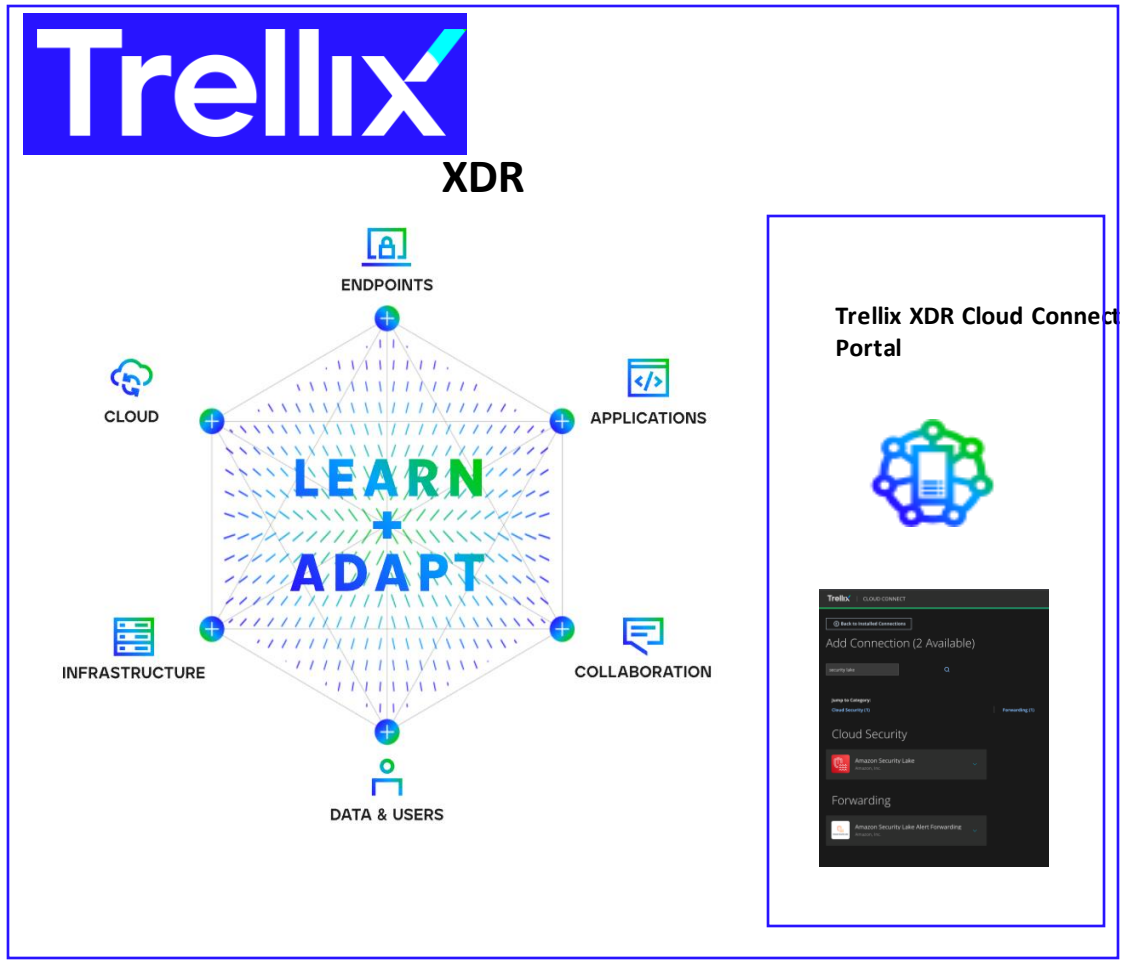


Figure 1: Joint customers can share security events across Trellix XDR and with Amazon Security Lake, getting complete detection and response capabilities for their AWS environments.

Example with LLM02: Insecure output handling

Amazon Bedrock

Malcious prompt from web app:

"Create Javascript that will send the session ID to this URL..."

Amazon CloudWatch

Trellix XDR can tie AI activity together with cloud and other platform events.

Trellix

XDR Investigative Tips

Built-in Expert Investigation

TIMELINE | AUTOMATIONS | INVESTIGATIVE TIPS | INTEL | EVENTS 100 | AFFECTED ASSETS 1 | HISTORY | NOTES | OS CHANGES

Investigative Tips provide a series of "next steps" for investigating an alert. For FireEye-provided rules, these searches are generated by incident responders and intelligence analysts based on the data they would look for to determine if an alert is a true positive. These searches are not meant to be all-inclusive, but they are designed to provide a place to start.

Expand All Queries

Did any other rules fire for this role? (8h Time Offset) Search not yet run

Did any other rules fire for this IP? (8h Time Offset) Search not yet run

What other sources have accessed this role? (8h Time Offset) Search not yet run

What other CloudTrail actions are there for this role? (1h Time Offset)

srczone	action	srcipv4	srcisp	srccountry	Count
ec2.amazonaws.com	startinstances	114.216.106.130	amazon technologies inc.	united states	200

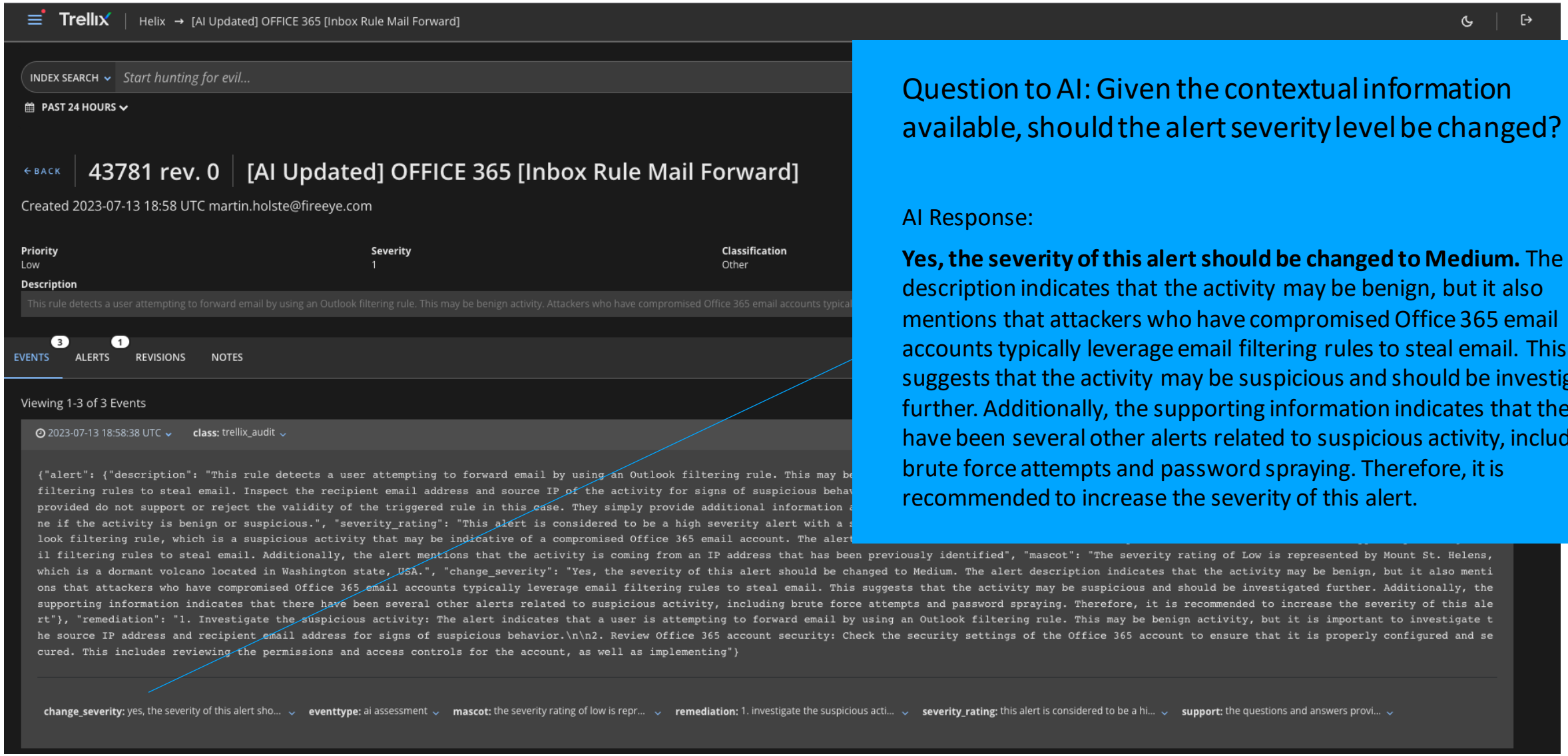
What other CloudTrail actions are there for this user? (1h Time Offset) Search not yet run

What other CloudTrail actions are there for this API key (if found)? (1h Time Offset) Search not yet run

What other CloudTrail events are there for this IP? (1h Time Offset)

srczone	action	srcipv4	srcisp	srccountry	Count
ec2.amazonaws.com	startinstances	114.216.106.130	amazon technologies inc.	united states	200
s3.amazonaws.com	putbucketacl	114.216.106.130	amazon technologies inc.	united states	2

Confidential –Do not distribute



Question to AI: Given the contextual information available, should the alert severity level be changed?

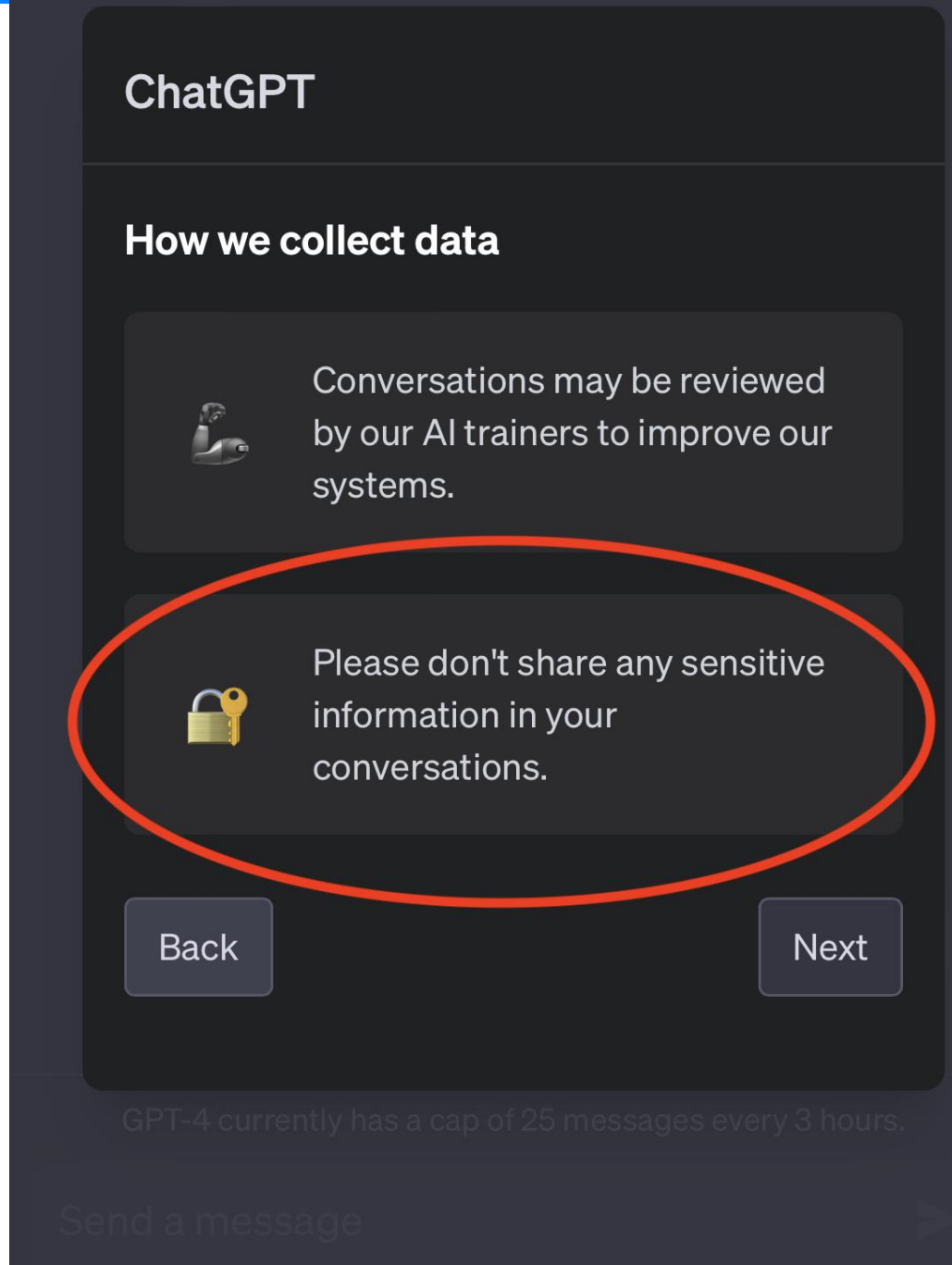
AI Response:

Yes, the severity of this alert should be changed to Medium. The alert description indicates that the activity may be benign, but it also mentions that attackers who have compromised Office 365 email accounts typically leverage email filtering rules to steal email. This suggests that the activity may be suspicious and should be investigated further. Additionally, the supporting information indicates that there have been several other alerts related to suspicious activity, including brute force attempts and password spraying. Therefore, it is recommended to increase the severity of this alert.

Trellix

What is the Trellix advantage in the XDR AI race?

1. We have more integrations than anyone else which gives us **the best data**.
2. We have hundreds of investigative tips and Infoseeker queries so we know **the right questions** to ask.
3. We operate **our own LLM's** so we don't send sensitive data to third parties or share between tenants.



Confidential –Do not distribute

Trellix XDR Generative AI Private Preview

The new Trellix XDR Generative AI Private Preview Program is a joint effort between Trellix and its strategic partners to explore ways of leveraging the rich data and investigation playbooks in Trellix XDR with Trellix-operated generative AI. The goal is to safely, privately, and automatically conduct as much investigation as possible to identify themes and insights that scripted playbooks and statistical analysis can miss.

- Trellix does this by providing a platform and framework for detection content within XDR for asking questions and processing answers from our hosted large language models (LLM's).

- All data is kept within the Trellix cloud environment, ensuring that no data is ever shared with other tenants or sent to a third party.

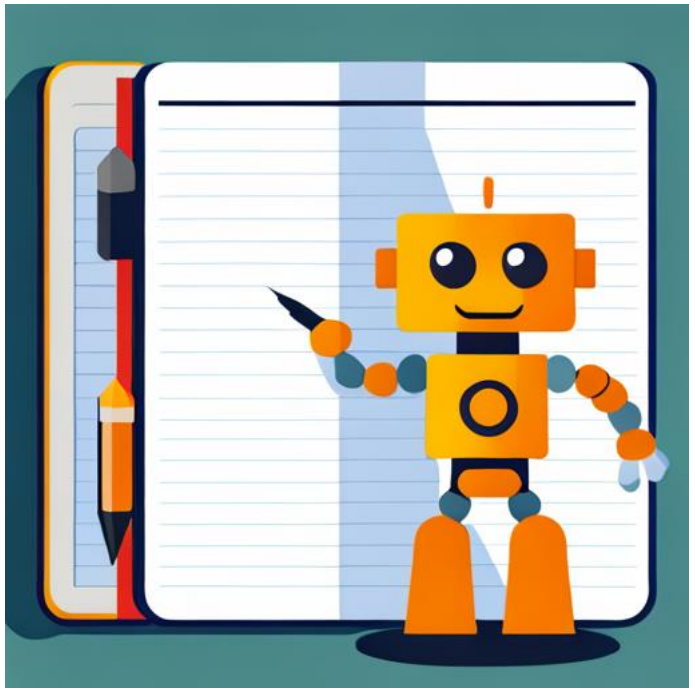
- The program is designed to foster collaboration between Trellix and our partners to build and refine the detection content so that Trellix XDR users are more efficient and surface findings that would otherwise have been missed.

- The program is considered experimental, and Trellix makes no guarantees regarding availability or implementation. However, we believe the collaboration is valuable and the experience will benefit those accepted into the program.

Trellix

Key Takeaways

1. The two main types of AI are machine learning and generative AI.
2. Generative AI is new and powerful, and it enables many opportunities.
3. Generative AI has many costs and considerations, and it warrants protection and oversight.



Trellix

Thank You

Trellix