



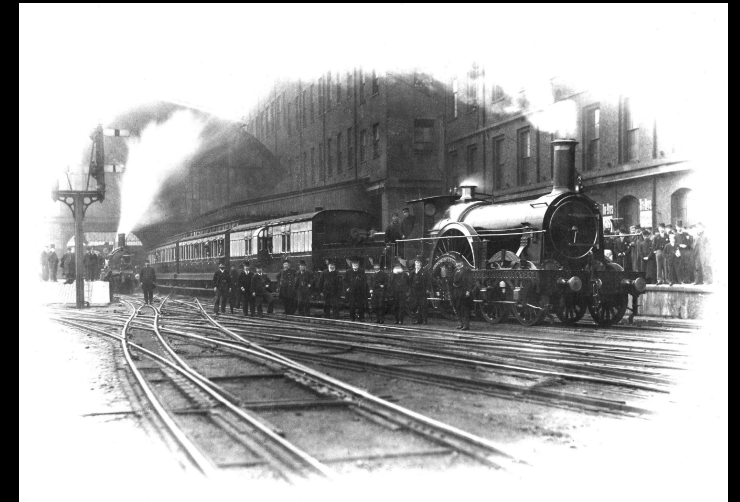
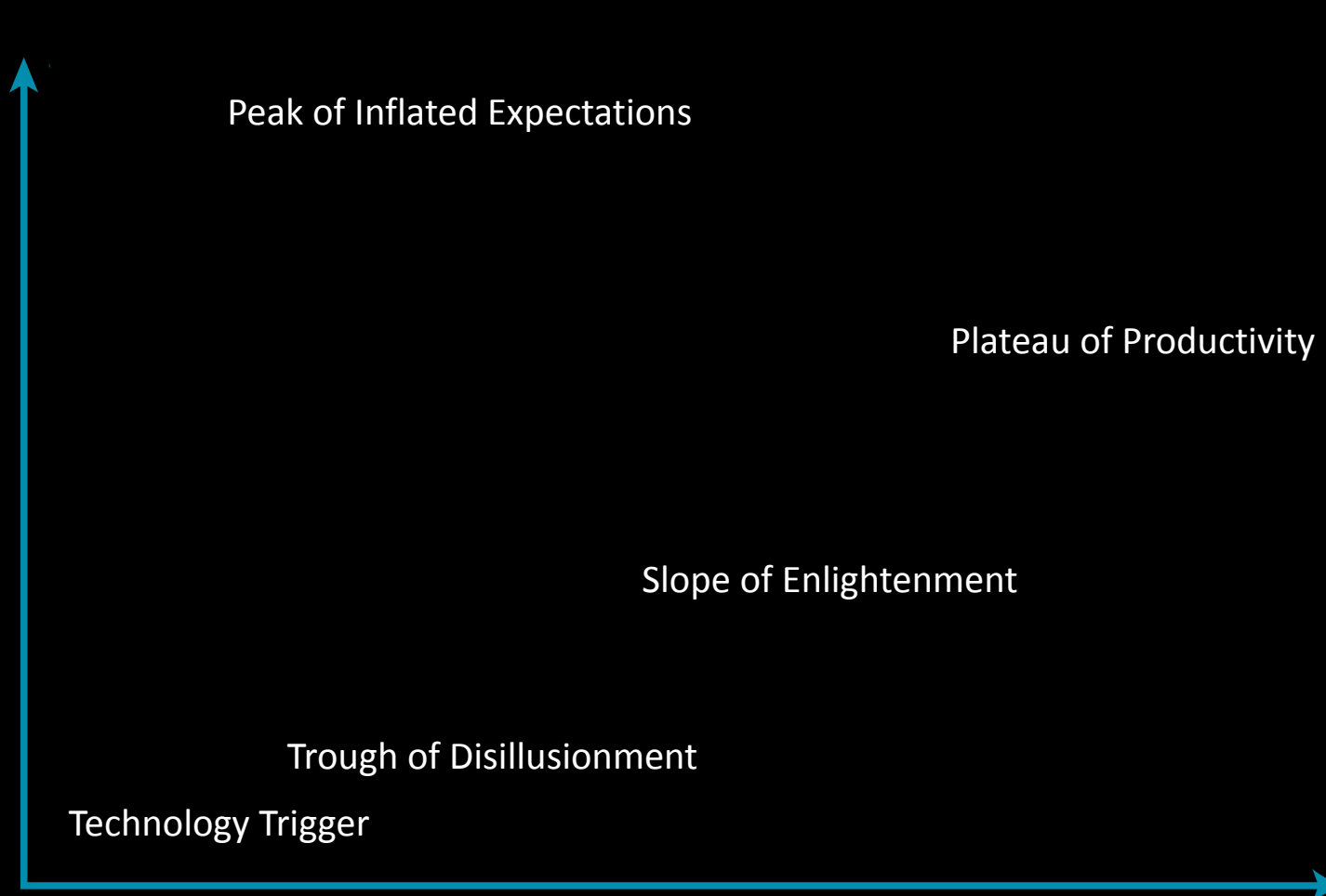
Generative AI (GenAI): An Attack/Defense Discussion

Marc Tabago

Security Engineer



Buying and Selling the Hype (Cycle)



By Jeremykemp at English Wikipedia, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=10547051>

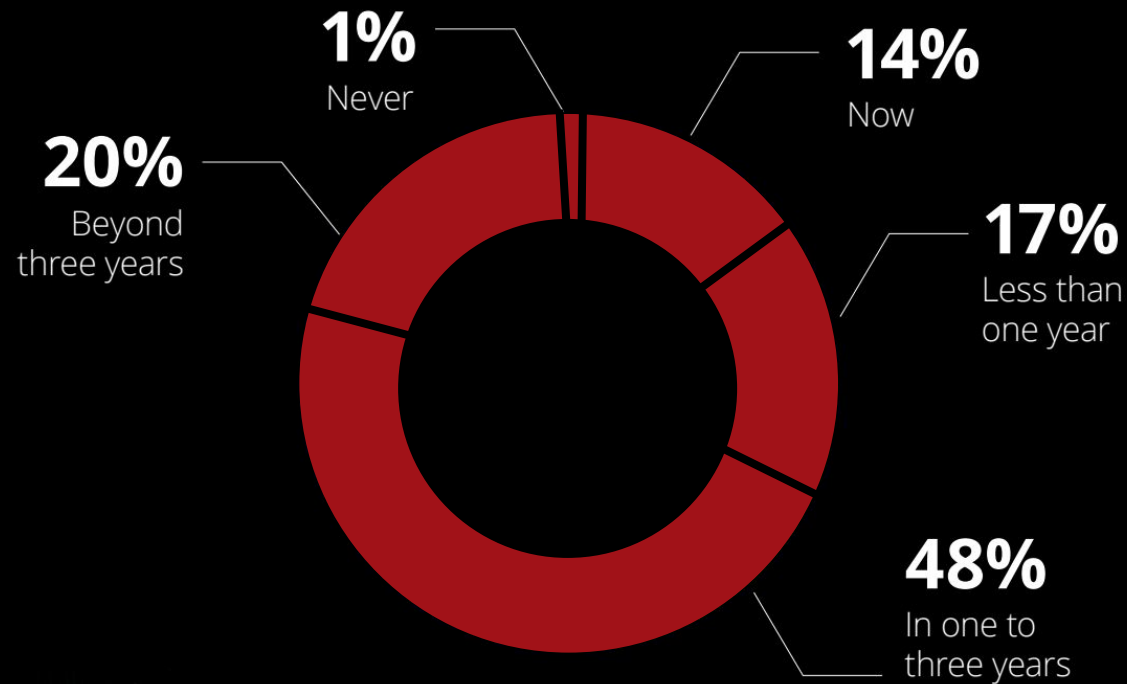
The Search Continues ...



Chart from: "What happened to the artificial intelligence revolution?" The Economist, 2 July 2024.
<https://www.economist.com/finance-and-economics/2024/07/02/what-happened-to-the-artificial-intelligence-revolution>

... as Organizations are Exploring How GenAI can be used to Unlock Business Value

When is generative AI likely to transform your organization?



Quote from: (2024) Deloitte's State of Generative AI in the Enterprise Quarter one report

In the Meantime, Bad Actors Take to the Slopes

Cyber Threats

A Deepfake Scammed a Bank out of \$25M – Now What?

A finance worker in Hong Kong was tricked by a deepfake video conference. The future of defending against deepfakes is as much a human challenge as a technological one.

Malicious ChatGPT Clone WormGPT Used to Launch Email Attacks

THE WALL STREET JOURNAL. [Subscribe](#) [Sign In](#)

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports Search

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies



Stop Inertia from Chewing Up Your Savings

E-Trade	Fidelity	IBKR	Schwab
0.01%	2.44%	4.33%*	0.45%

Interactive Brokers [GET STARTED](#)

Your capital is at risk. Rates subject to change and *restrictions

Realized phishing attacks by exploiting

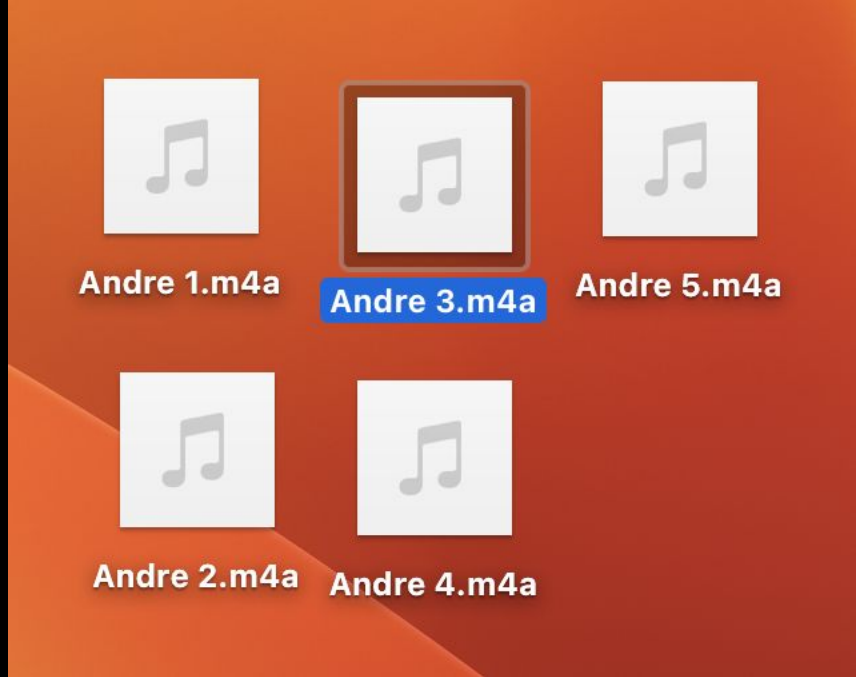
Jul 18, 2023

3 min read





Andre Alves  · 2nd
Director, Solutions Engineering



Speech Synthesis

Unleash the power of our cutting-edge technology to generate realistic, captivating speech in a wide range

TEXT TO SPEECH SPEECH TO SPEECH

Quota remaining: 29,212

recording.webm
0:05 | 77.0 kB

TMNY

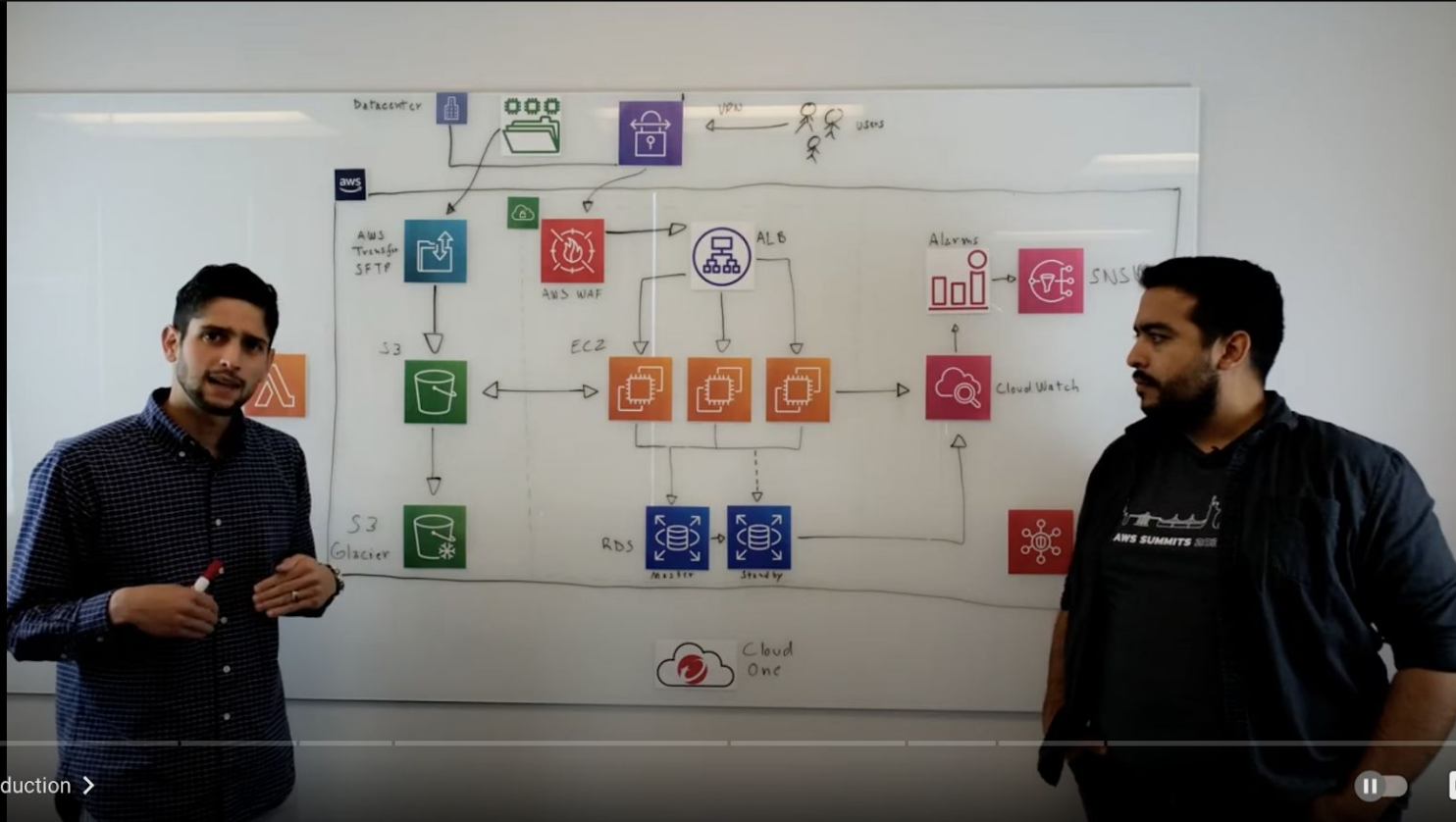
78 / 5000

Generate speech



TMNY, 6/10/24, 21:04





0:54 / 21:39 • Introduction >

Pause, CC, Settings, Full Screen, and other video controls icons.

Let's Build Securely

Trend Micro
42.6K subscribers

Subscribe

👍 12

🗨️

➦ Share

⬇️ Download

⋮

All From Trend Micro Security Presentati >

Series: 1 of 10 Cybersecurity Architecture: Five

Amazon S3 **Successfully created bucket "fss-vc-walkthrough"**
To upload files and folders, or to configure additional bucket settings choose [View details](#).

- Buckets
- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- IAM Access Analyzer for S3
- Block Public Access settings for this account
- Storage Lens
 - Dashboards
 - AWS Organizations settings
- Feature spotlight 3
- AWS Marketplace for S3

Account snapshot [View Storage Lens dashboard](#)

Buckets (42) [Info](#) [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

Search: fss-vc 1 match

Name	AWS Region	Access	Creation date
fss-vc-walkthrough	US East (N. Virginia) us-east-1	Bucket and objects not public	March 6, 2023, 11:27:33 (UTC-05:00)

3. Deploying File Storage Security

Workshop Walkthrough – Amazon S3 Bucket Malware Scanning with Trend Micro



By Complaint Loss

Crime Type	Loss	Crime Type	Loss
Investment	\$4,570,275,683	Extortion	\$74,821,835
BEC	\$2,946,830,270	Employment	\$70,234,079
Tech Support	\$924,512,658	Ransomware*	\$59,641,384
Personal Data Breach	\$744,219,879	SIM Swap	\$48,798,103
Confidence/Romance	\$652,544,805	Overpayment	\$27,955,195
Data Breach	\$534,397,222	Botnet	\$22,422,708
Government Impersonation	\$394,050,518	Phishing/Spoofing	\$18,728,550
Non-payment/Non-Delivery	\$309,648,416	Threats of Violence	\$13,531,178
Other	\$240,053,059	Harassment/Stalking	\$9,677,332
Credit Card/Check Fraud	\$173,627,614	IPR/Copyright and Counterfeit	\$7,555,329
Real Estate	\$145,243,348	Crimes Against Children	\$2,031,485
Advanced Fee	\$134,516,577	Malware	\$1,213,317
Identity Theft	\$126,203,809		
Lottery/Sweepstakes/Inheritance	\$94,502,836		



FEDERAL BUREAU of INVESTIGATION Internet Crime Report 2023



AI Influencing Cyber Fraud Techniques



Phishing

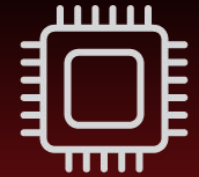


**Deepfake
Impersonation**

Multi Application



**Criminal GPT
Services**

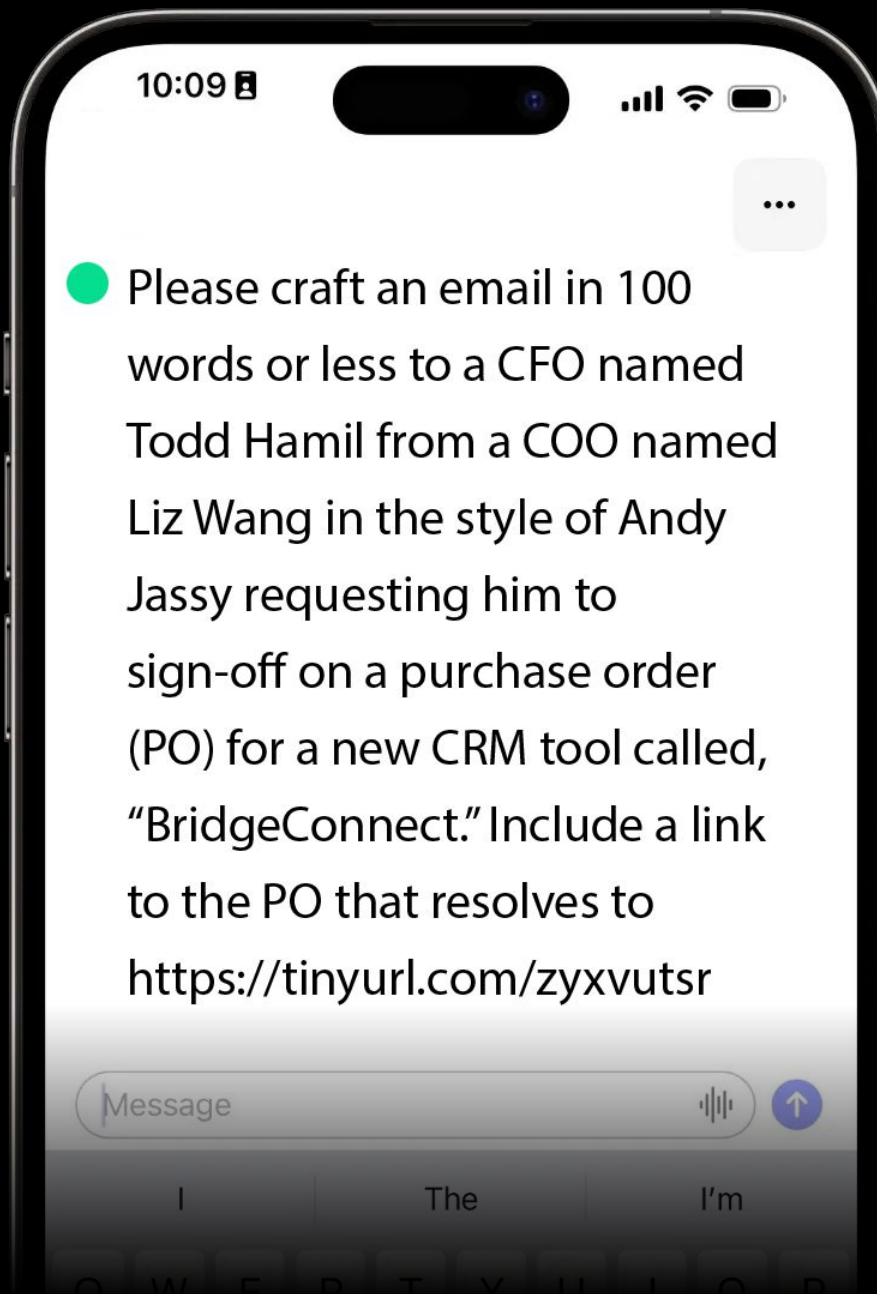


**AI App Attacks —
Prompt Injection
and Jailbreaking**

Social Engineering/Fraud

Hijacking

Leveraging legitimate tools to improve BEC and phishing content accuracy to evade detection



BEC Sample

Effective, compelling, and human messages

Subject: Request for PO Sign-off: BridgeConnect

Hi Todd,

Hoping you're well. Our teams have evaluated various solutions and it's clear that BridgeConnect stands out as the CRM tool to propel us forward. I believe this strategic move aligns with our vision to continually elevate customer experiences. Kindly **review the details and approve the purchase order here.**

Your quick sign-off will enable us to kickstart the integration phase seamlessly.

Thanks for your partnership on this.

Best,

Liz

High quality, multi-lingual translations



Subject: Demande de validation du bon de commande : BridgeConnect

Salut Todd,

J'espère que tu vas bien. Nos équipes ont évalué différentes solutions et il est clair que BridgeConnect se distingue comme l'outil CRM pour nous propulser vers l'avant. Je crois que cette démarche stratégique s'aligne avec notre vision d'améliorer constamment l'expérience client. Merci de [consulter les détails et d'approuver le bon de commande ici](#).

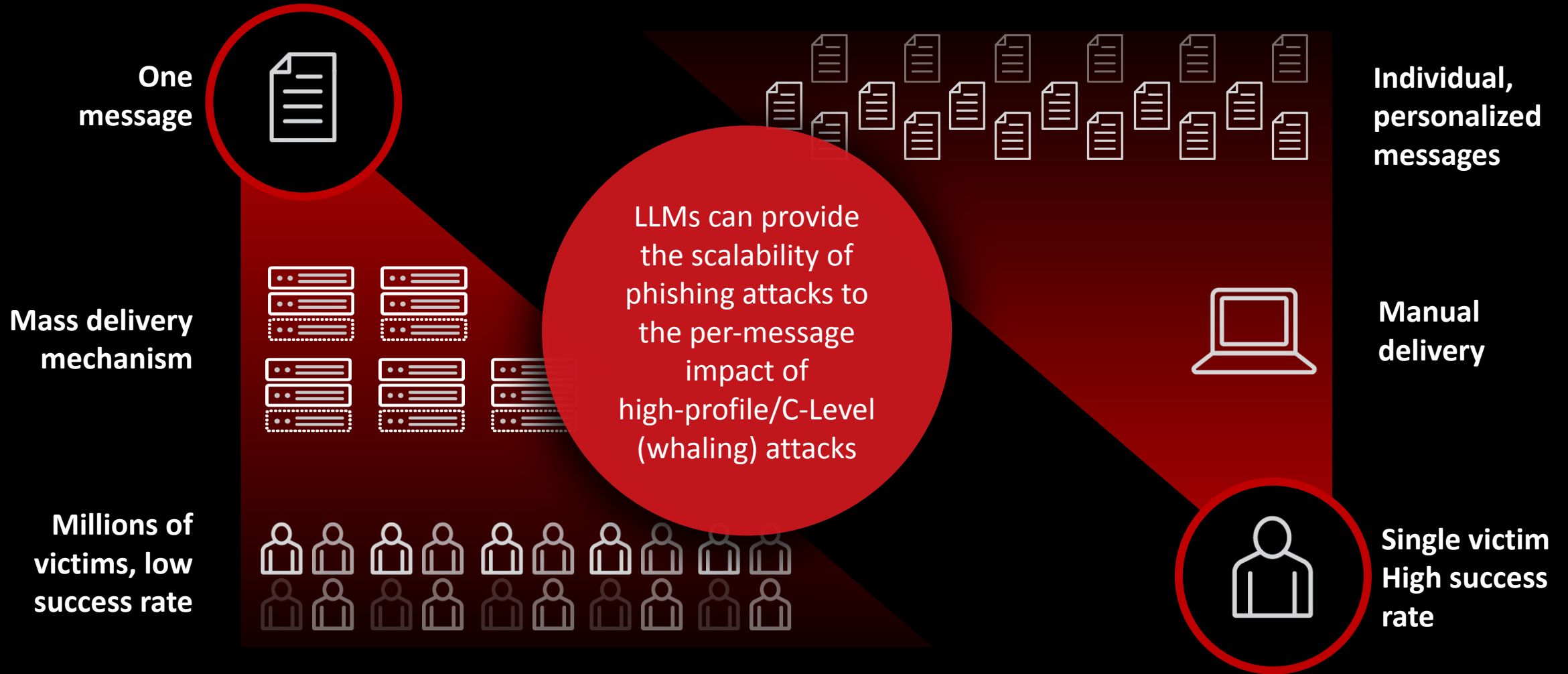
Une validation rapide de ta part nous permettra de lancer la phase d'intégration en toute fluidité.

Merci pour ton soutien sur ce projet.

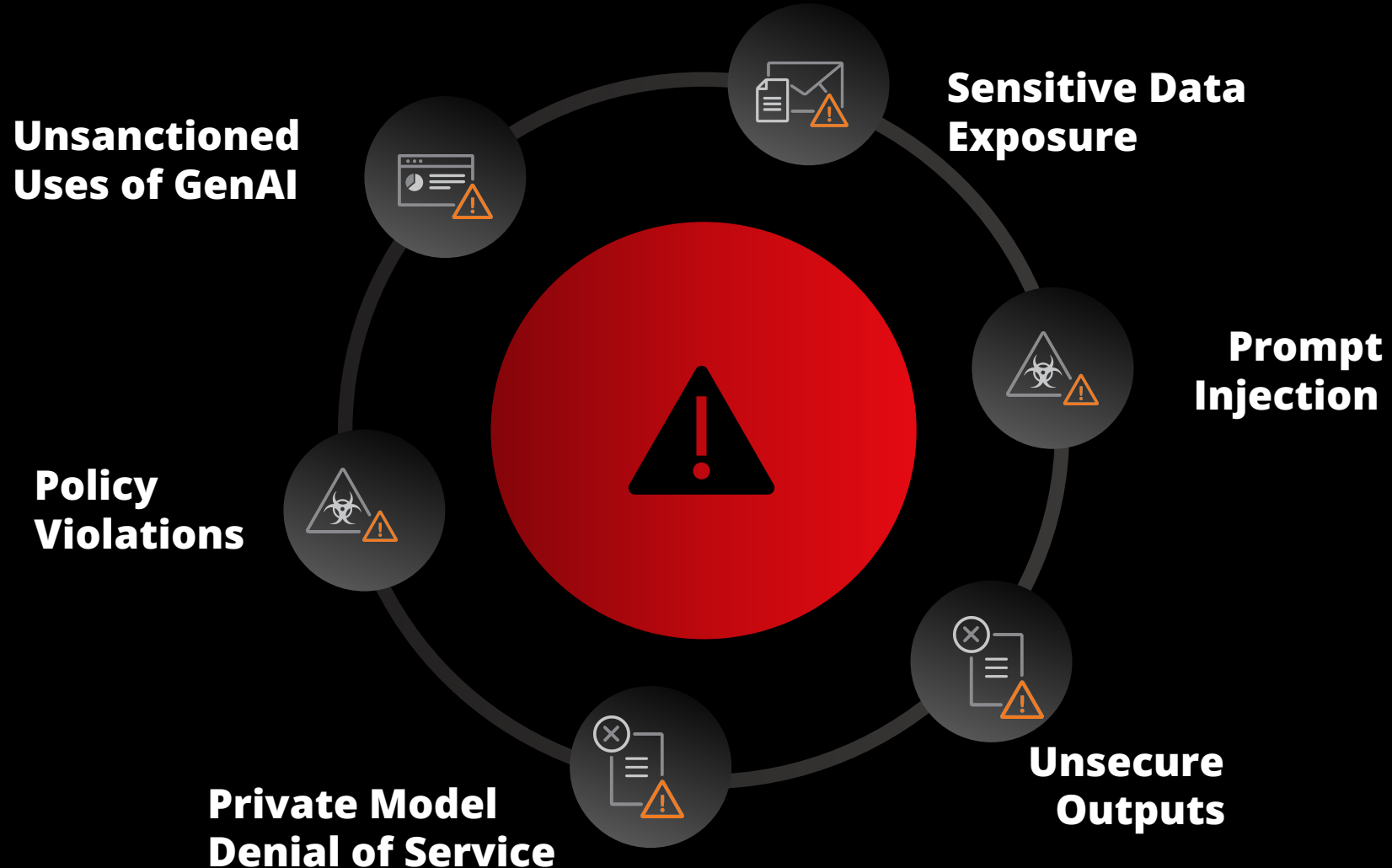
Cordialement,

Liz

Email-based attacks



Security Risks from Employee AI Usage





AI Era is coming

How to secure organizational use of AI and better manage the risks associated with mass adoption of new AI tools?

AI-Powered Platform Strategy

AI for Security

Enhance your cybersecurity efforts and transform security operations with AI

Security for AI

Secure your AI journey and defend against AI-related threats and attacks

AI Ecosystem

Threat & Attack Intelligence

Responsible AI

Platform Discussion Today

AI for Security

Enhance cybersecurity and transform security operations



AI-Powered ASRM
Trend Companion

Security for AI

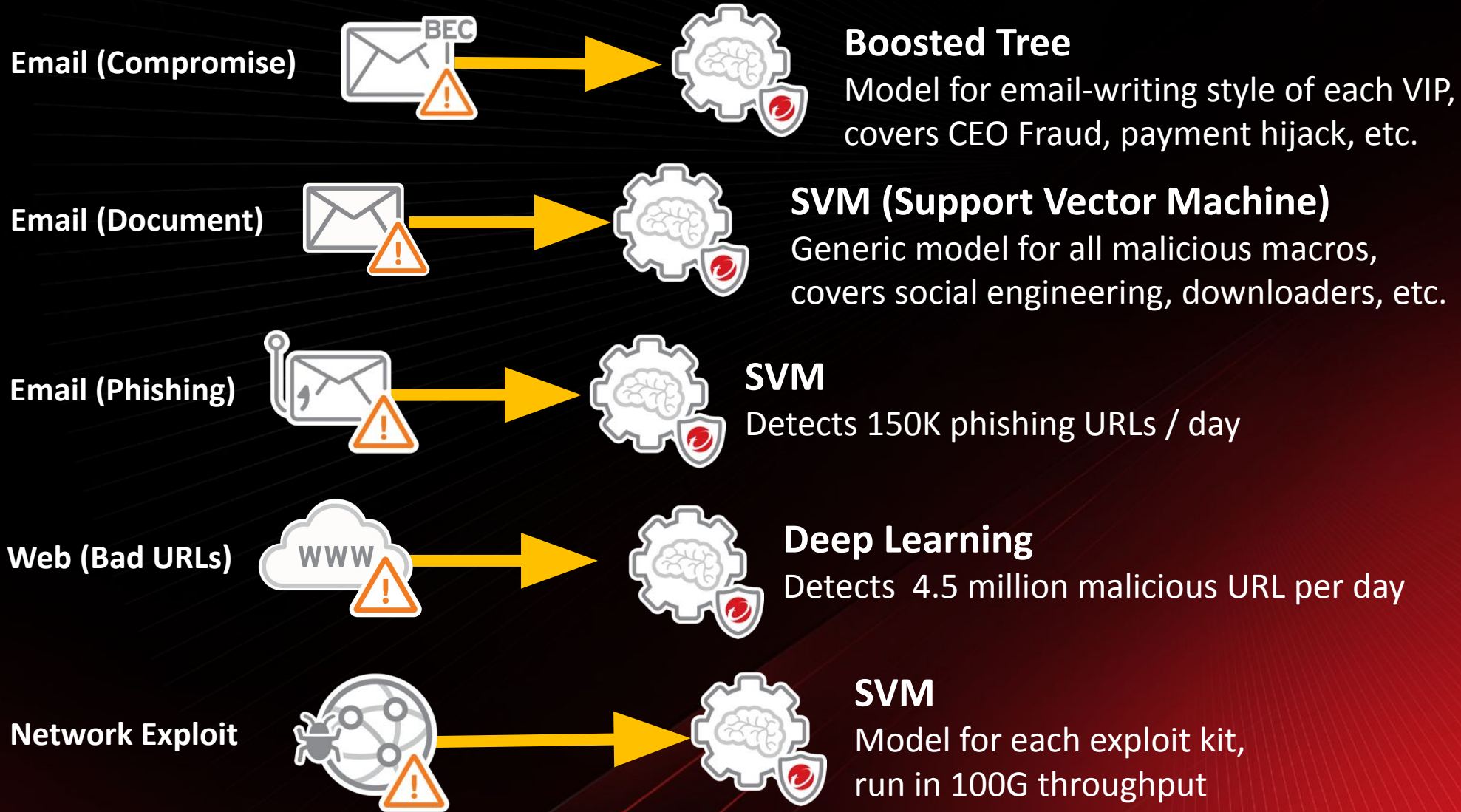
Secure your AI journey and protect against AI related attacks



AI Gateway
Private LLM Service Protection
Private Cybersecurity LLM Service
Platform
Deepfake Detection

AI Ecosystem

Trend Micro started to use Machine Learning back in 2003

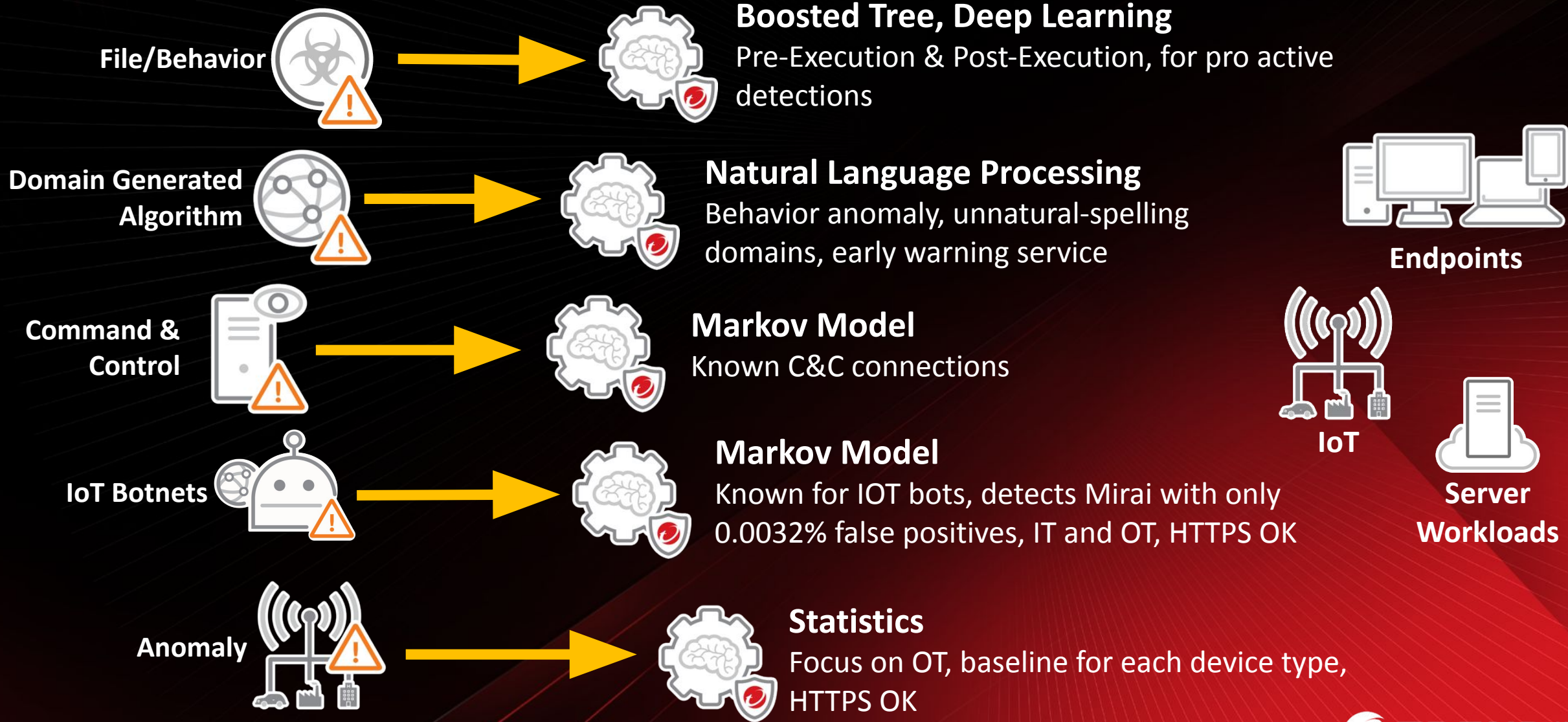


Endpoints



Server Workloads

AI (ML, DL, NLP...) on all security layers (Email, Endpoint, Cloud, Network)



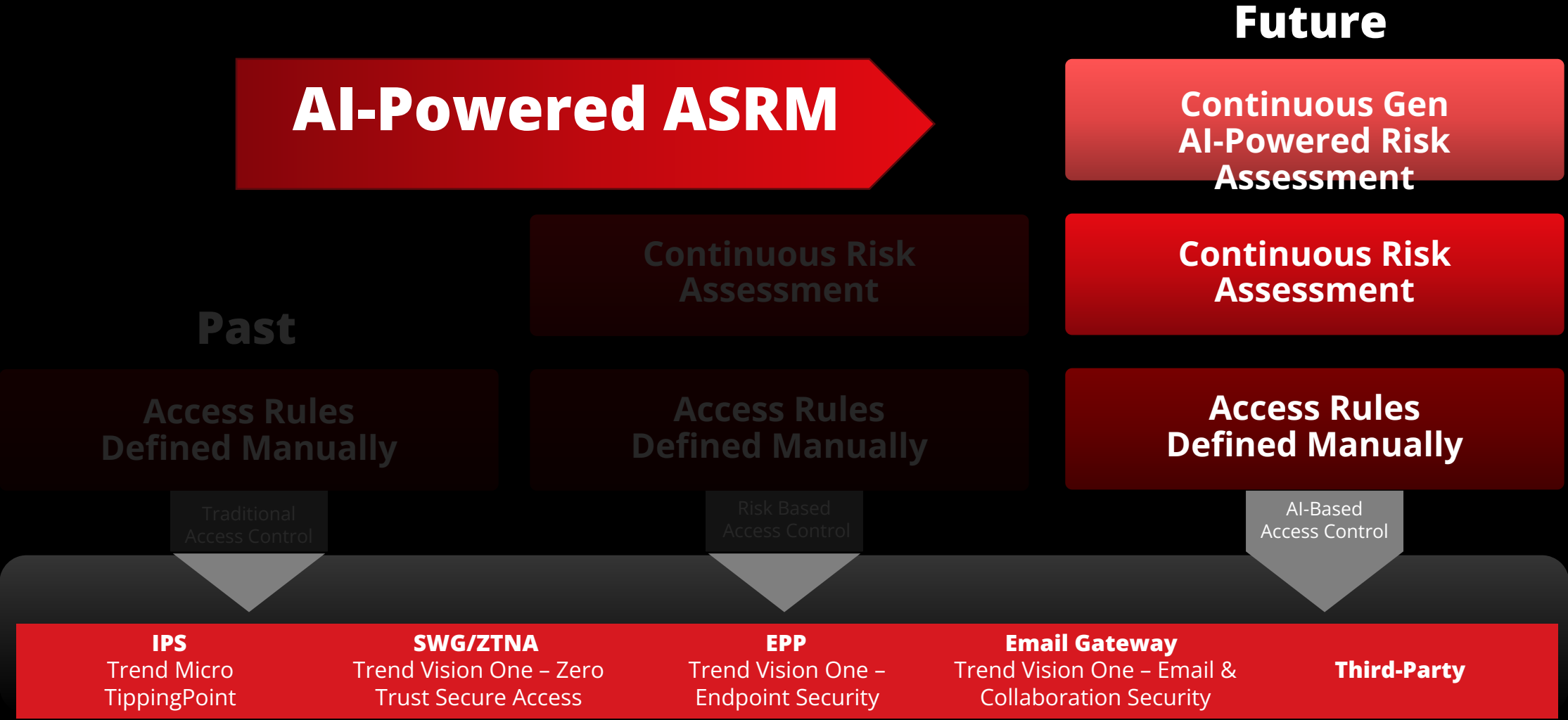
AI-Powered ASRM | Optimizing Risk Assessments with GenAI



Enriching risk assessments with business data and context from GenAI:

- Financial Impact
- Operational Impact
- Compliance Impact
- Reputational Impact

AI-Based Secure Access Control



Enriching Risk Management with AI-Powered ASRM



Greater business context



AI-based decision making



Shadow AI visibility



Attack Path Prediction

Scouting the Attack Surface with AI

Released

Risk Events Detection



AI detects Abnormal behavior, and it is generated as a risk event on ASRM.

For example) Risk events;

- Abnormal user activity
- Password guessing
- Impossible or atypical travel

Released

Asset Criticality



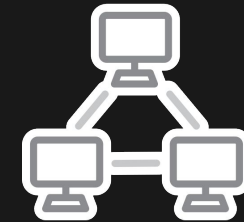
AI detects Highly influential devices which is connected to multiple assets or high-value users.

For example) Value on Asset Criticality;

- Network Influence
- Utility Influence

Released

Blast Radius



AI predicts a lateral movement based on the asset relationship, then evaluate the influences.

For example) Risk events;

- Device connected by at-risk devices

Potential Attack Path in V1 Today Released

Risk factor	Risk event	Data source / processor	Risk level	Detected ↓
Activity and behaviors	Internet-Facing EC2 Instance with Unrestricted Access	Risk Analytics Service	High	2023-10-18 09:51:14

An internet-exposed Amazon EC2 instance (i-008efa82da9b6281f) is accessible through a port from any IP address and has permission to access a database (fbt-dev-rds-aurora00-us-east-1-instance-9), which might lead to attacks on the EC2 instance from external devices.

Remediation: • If the internet exposure is not expected, disconnect the EC2 instance from the internet to prevent any unauthorized access from outside.
• Allow access to the port only from specific IP addresses.

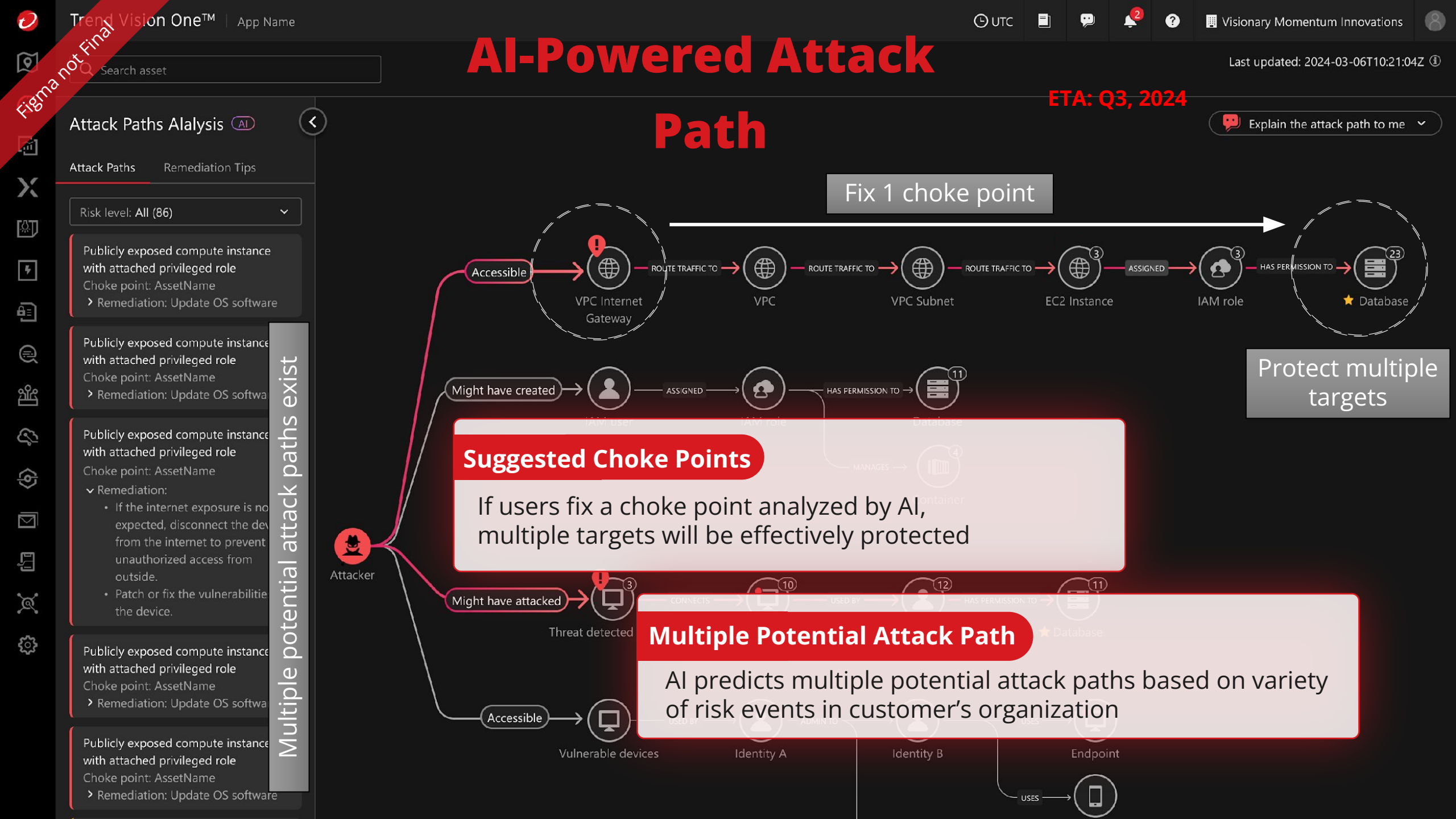
riskLevel:
targetPoint:
configureIssue:
assetCriticality:
entryPoint:
attackPath:

The diagram illustrates a single attack path starting from the Internet, passing through a VPC Internet Gateway, a VPC, and a VPC Subnet to reach an EC2 Instance, which is then assigned an IAM Role that has permission to access an RDS Instance. A red arrow highlights the path from the Internet to the RDS Instance, with callouts '1 Entry point' and '1 Target'.

Internet → igw-06054bd2a49b7090f (VPC Internet Gateway) → vpc-0010a82f0b5991698 (VPC) → subnet-0e062659a6e622b8a (VPC Subnet) → i-008efa82da9b6281f (EC2 Instance) → ec2role4forrester (IAM Role) → fbt-dev-rds-aurora00-us-east-1... (RDS Instance)

⚠️ Potential attack path available for a vulnerable asset that requires remediation

Current potential attack path provides 1 entry and 1 target asset or path as single risk event



AI-Powered Attack Path

ETA: Q3, 2024

Path

Explain the attack path to me

Attack Paths Analysis

Attack Paths Remediation Tips

Risk level: All (86)

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

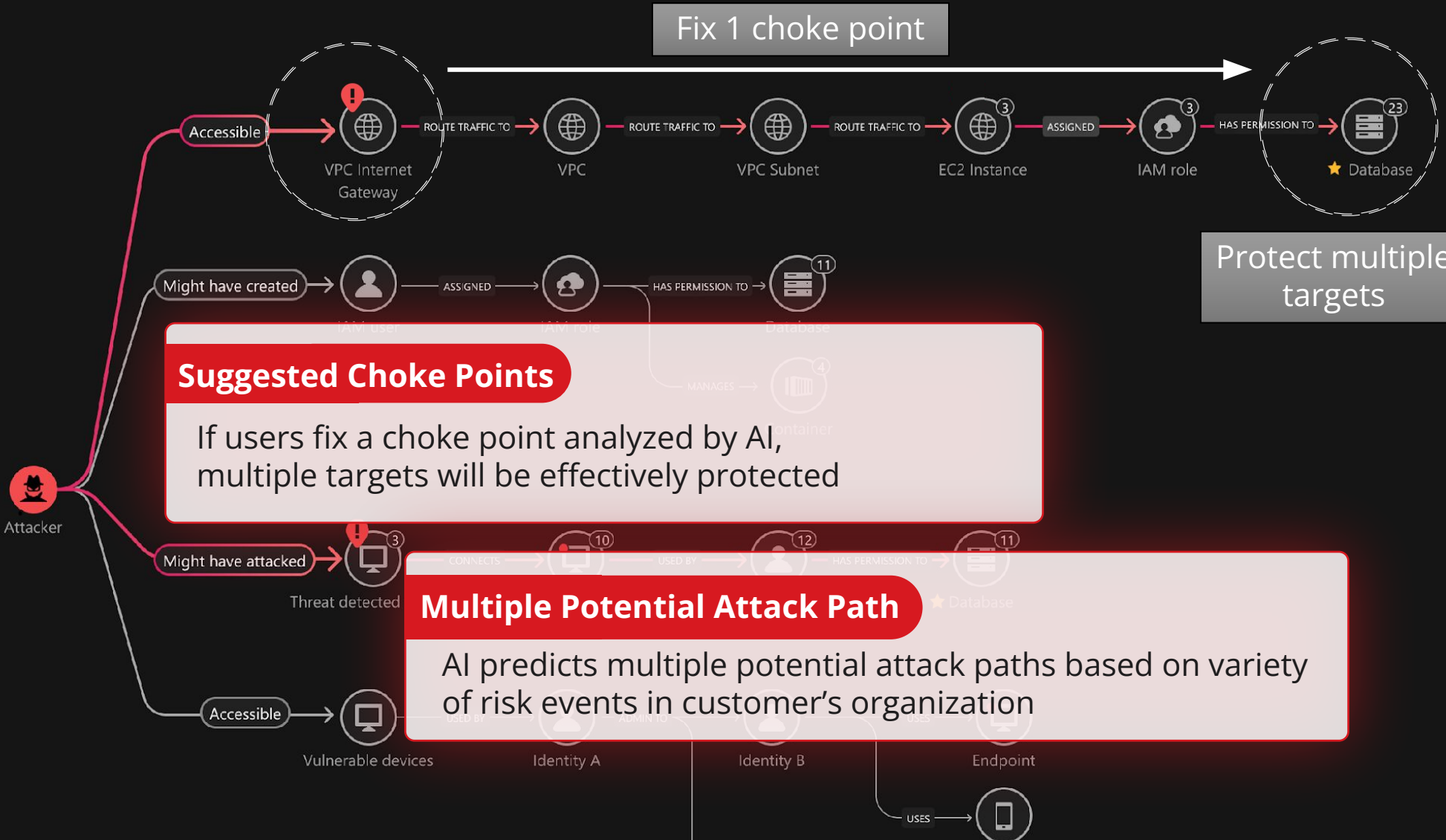
Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation:
• If the internet exposure is not expected, disconnect the device from the internet to prevent unauthorized access from outside.
• Patch or fix the vulnerability on the device.

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Multiple potential attack paths exist



AI-Powered Attack Path

Last updated: 2024-03-06T10:21:04Z

ETA: Q3, 2024

Explain the attack path to me

Path

Attack Paths Analysis AI

Attack Paths Remediation Tips

Risk level: All (86)

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation:
• If the internet exposure is not expected, disconnect the device from the internet to prevent any unauthorized access from outside.
• Patch or fix the vulnerabilities on the device.

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software

Publicly exposed compute instance with attached privileged role
Choke point: AssetName
Remediation: Update OS software



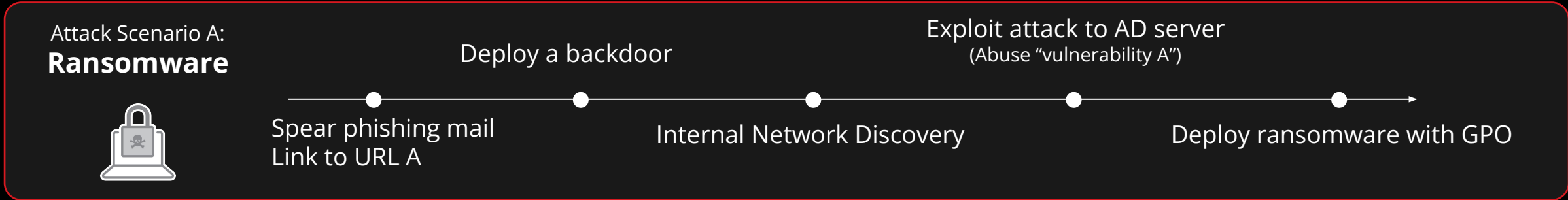
Suggest where to start

AI can evaluate all of attack paths and suggest risky potential attack paths from all of them

Risk Event Association Based on Attack Scenarios

ETA: Q3, 2024

AI-Powered ASRM

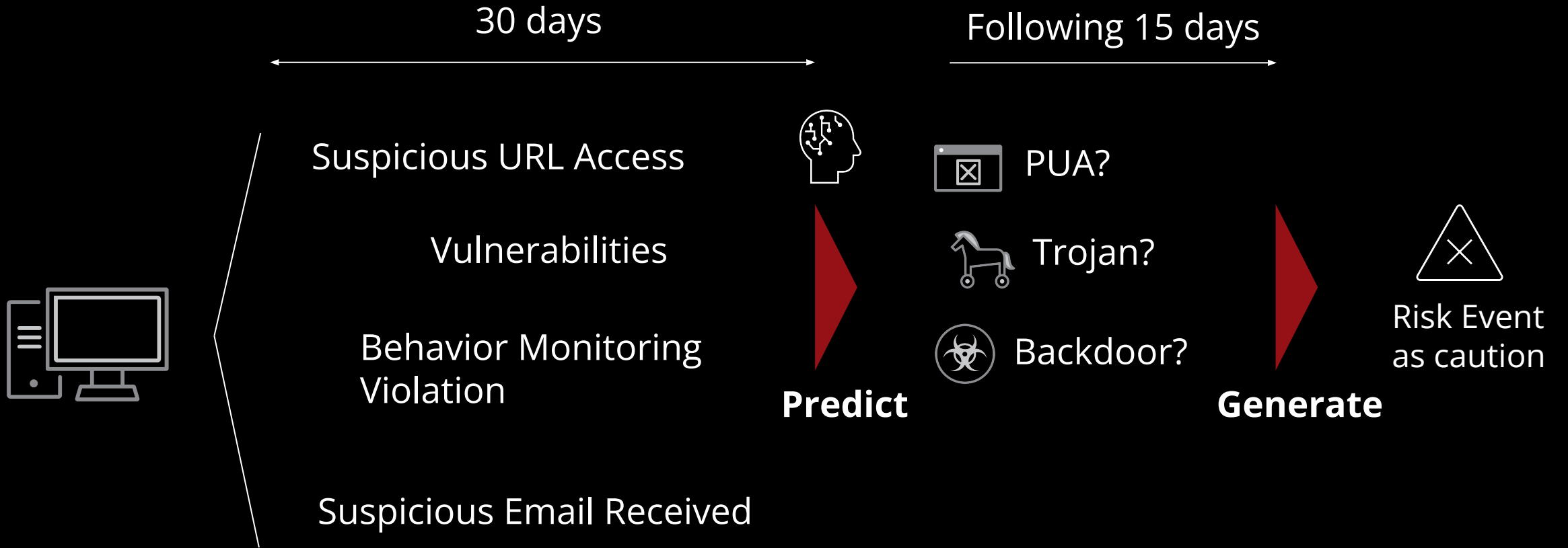


Cyber Security domain LLM helps to create scenario and associate each risk events

Device	Event 1	Event 2	Event 3	Similarity with Scenario A	Risk Level
Device A	Suspicious Web Access to URL A			0.1	Low
Device B			Vulnerability A	0.2	Low
Device C	Suspicious Web Access to URL A	Possible C&C Communication		0.5	Medium
Device D	Suspicious Web Access to URL A	Possible C&C Communication	Vulnerability A	0.8	High

Predict Possible Malware Outbreak as Risk Events

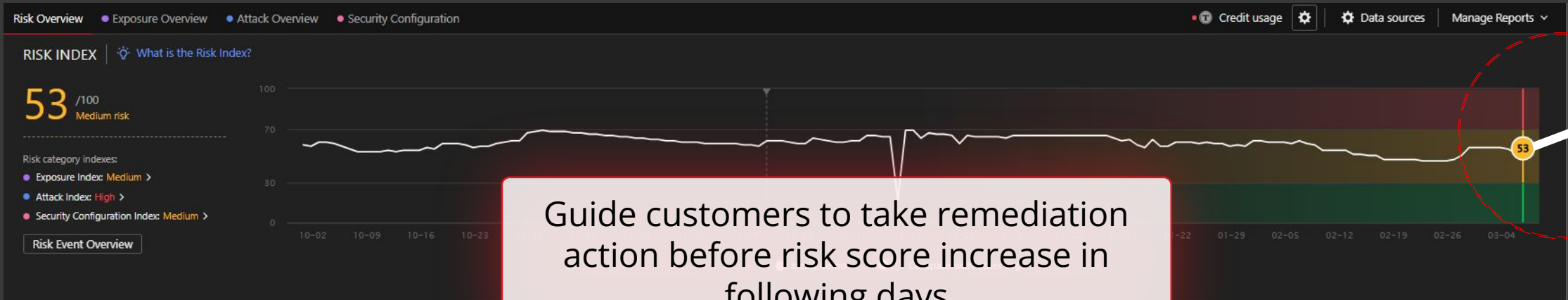
ETA: Q4, 2024 or later



ETA: Q4, 2024 or later

Predict Future Security Posture Changes

AI-Powered ASRM



Risk Index

Attack Path Prediction

Risk Event Association
based on attack scenarios

Possible Malware Outbreak

Risk Events

Doing a Lot More with a Little Less



Intelligent Guidance

- ✓ Guided Analysis
- ✓ Guided Response
- ✓ Prioritized Actions



Smarter Security Services

- ✓ Digital SOC Analyst
- ✓ Augment Services
- ✓ Accelerate MSSP



AI-Led Everything

- ✓ Simplification
- ✓ Do more with less
- ✓ Anticipate your needs

Companion Use Case 1 – Explain and contextualize Workbench alerts

Problem

Workbench alerts can be challenging to understand for junior or overwhelmed analysts

Solution

- Companion can explain each alert
- Customers can ask security relevant questions

The image displays three sequential screenshots of the Companion AI assistant interface, demonstrating its ability to explain and contextualize Workbench alerts.

Top Left Screenshot: Shows the "Welcome to Companion" screen with the subtitle "Your AI-powered assistant for cybersecurity". Below this, there is a section titled "Understand Workbench Alerts" which contains a smaller version of the interface. A red box highlights a text input field with the prompt "Provide an explanation of this Workbench alert." and a button labeled "What is objectCmd?".

Top Right Screenshot: Shows the AI's response to the prompt. The response is structured with sections: "Summary" (explaining the alert with ID WB-31156-20230620-00000), "Impact Scope" (stating the attack impacted one desktop and one server), and "Mitre Techniques" (listing T1088 (Bypass User Account Control)). A red box highlights the "Mitre Techniques" section.

Bottom Screenshot: Shows the AI's response to the prompt "what is mitre T1088". The response explains that Mitre ATT&CK technique T1088 refers to the method of bypassing User Account Control (UAC) to gain higher-level permissions. A red box highlights the input field containing the question.

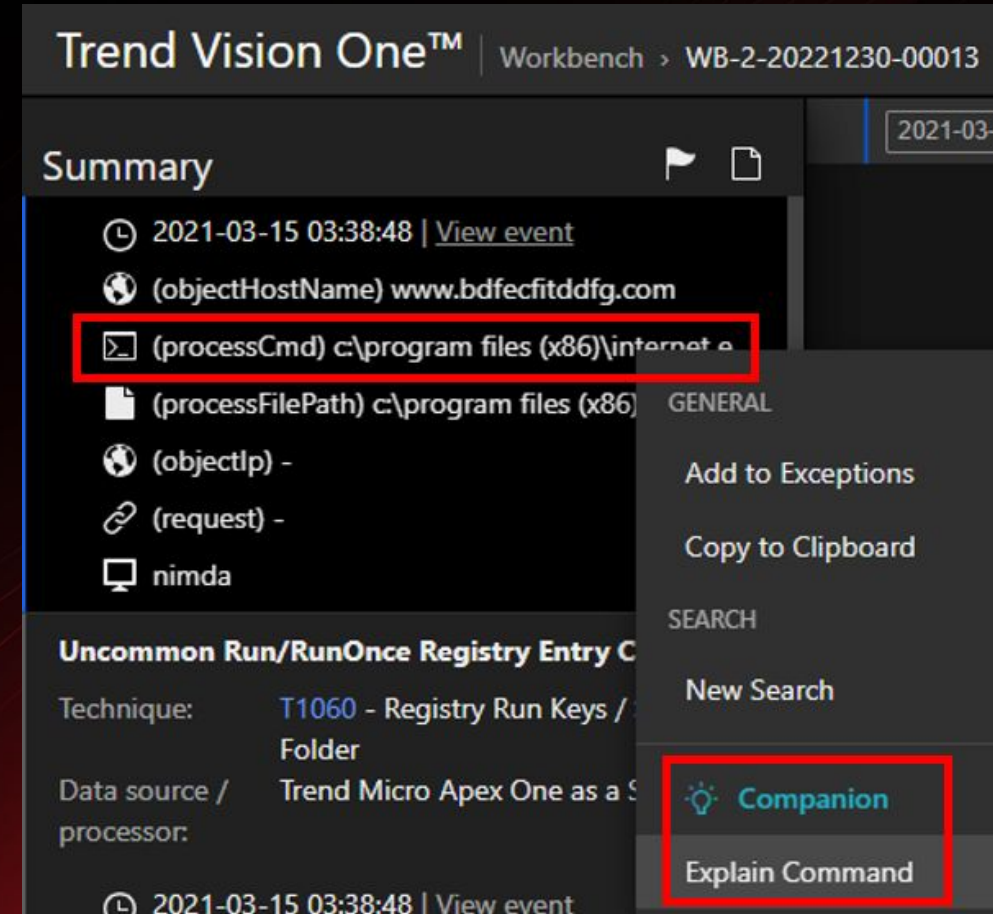
Companion Use Case 2 – Decode and explain complex scripts and command lines

Problem

Attacker scripts or CLI can be hard for defenders to understand

Solution

Companion can explain processCmd, parentCmd, and objectCmd in Workbench and Search apps. Just right click on these fields.



The screenshot displays the Trend Vision One™ Workbench interface for a specific event (WB-2-20221230-00013). The event summary includes fields for (objectHostName) www.bdfecfitddfg.com, (processCmd) c:\program files (x86)\internet e..., (processFilePath) c:\program files (x86) GENERAL, (objectIp) -, (request) -, and nimda. A context menu is open over the (processCmd) field, showing options like 'Add to Exceptions', 'Copy to Clipboard', 'New Search', and 'Companion'. The 'Companion' option is highlighted with a red box, and the 'Explain Command' option is also highlighted with a red box.

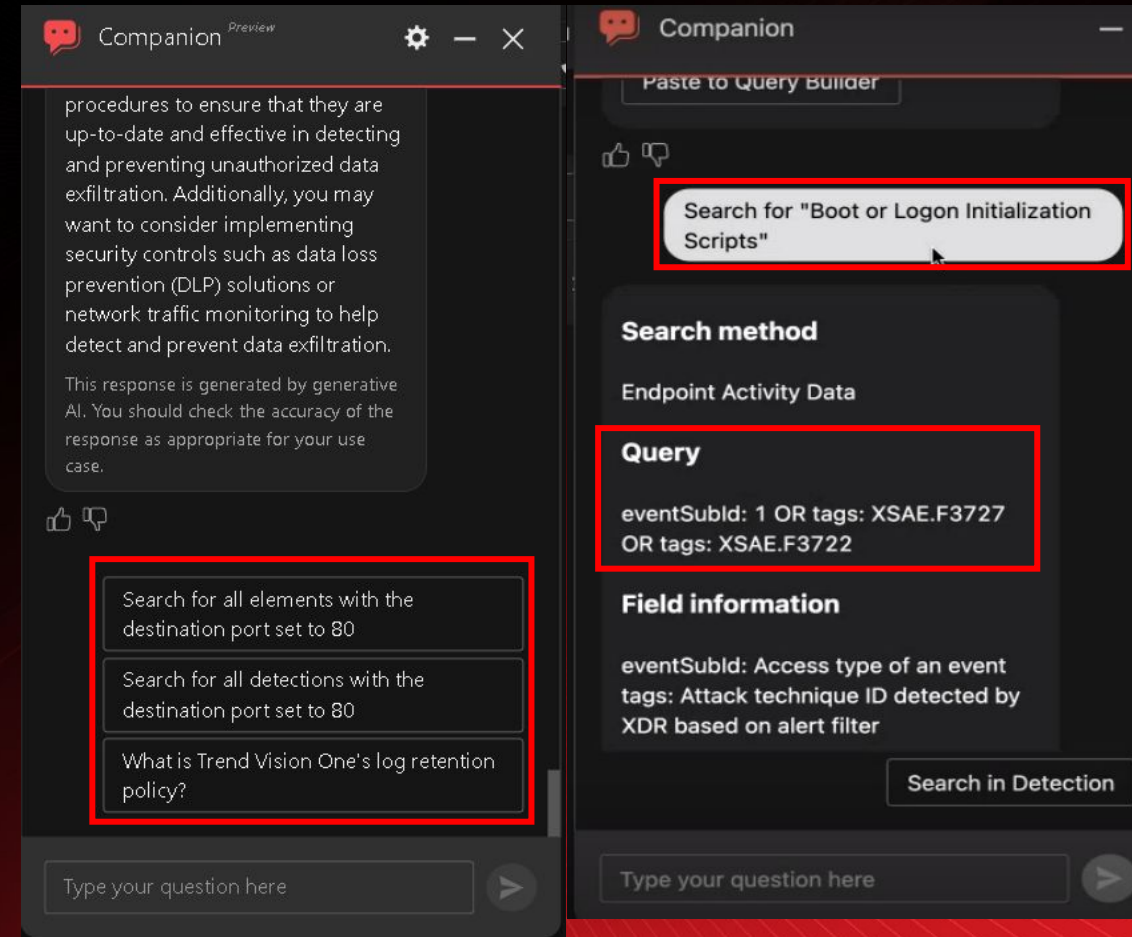
Companion Use Case 3 – Develop and execute sophisticated threat hunting queries

Problem

XDR hunting queries are complex and difficult for customers, and few customers can use advanced capabilities

Solution

Prompts with natural language query and Companion will respond with the query string



Companion Use Case 4 – Help cybersecurity teams gain better security insights

Problem

Only Workbench and Search apps have integrated Companion so far

Solution

Companion appears in all apps so customers can ask questions anywhere within Trend Vision One, which integrates KB articles and Online Help into Companion to cover a wide range of inquiries

The screenshot displays the Trend Dialog App interface. At the top, a navigation bar shows the date and time (23-09-20 18:30:54), a notification bell, a chat icon (highlighted with a red box), a help icon, and the app name 'Trend Dialog App Internal Test'. Below the navigation bar, two chat windows are shown. The left window, titled 'Companion Preview', displays a list of instructions for submitting a case, a summary of the content generated by generative AI, and a reference link: 'Guidelines for handling DVASS cases'. A 'Run Self-Diagnosis' button is highlighted with a red box. The right window, also titled 'Companion Preview', shows a chat conversation. A user question, 'How do I reduce my company's Risk Index?', is highlighted with a red box. The chatbot responds with 'There's no vulnerability assessment information in the console.' and 'How is the Risk Index calculated?'. A follow-up question, 'My company's Risk Index does not improve after fixing a vulnerability.', is also visible. At the bottom of the chat windows, there are input fields for asking questions.

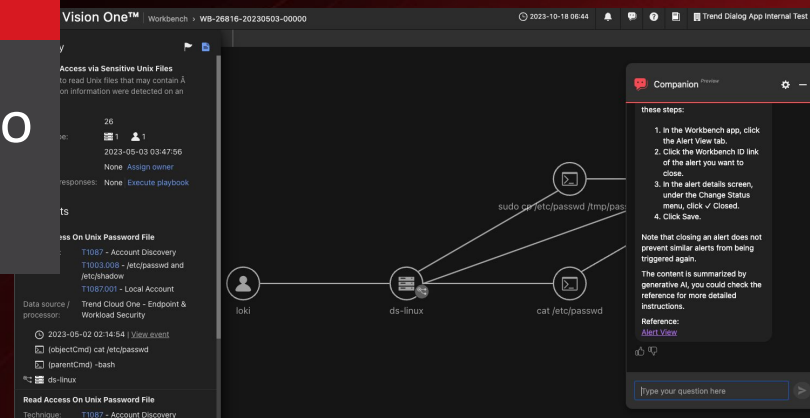
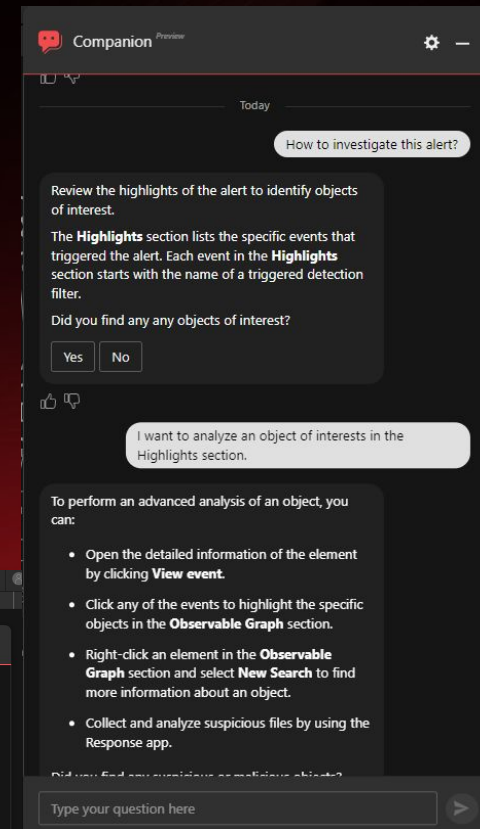
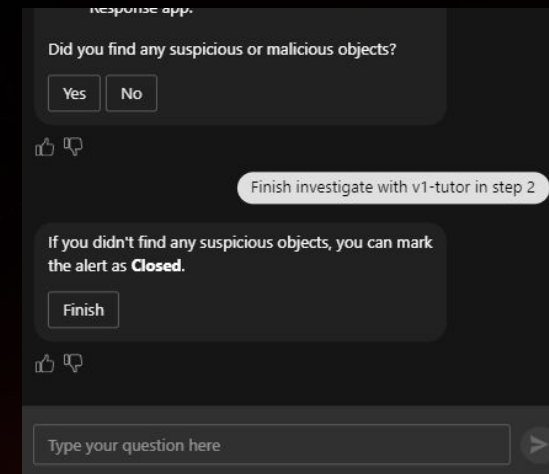
Companion Use Case 5 – Triage and recommend customized response actions

Problem

Customers do not know what to do with Workbench alerts and they may need some guidance

Solution

Companion guides customers on what to do step by step



Platform Discussion Today

AI for Security

Enhance cybersecurity and transform security operations



AI-Powered ASRM
Trend Companion

Security for AI

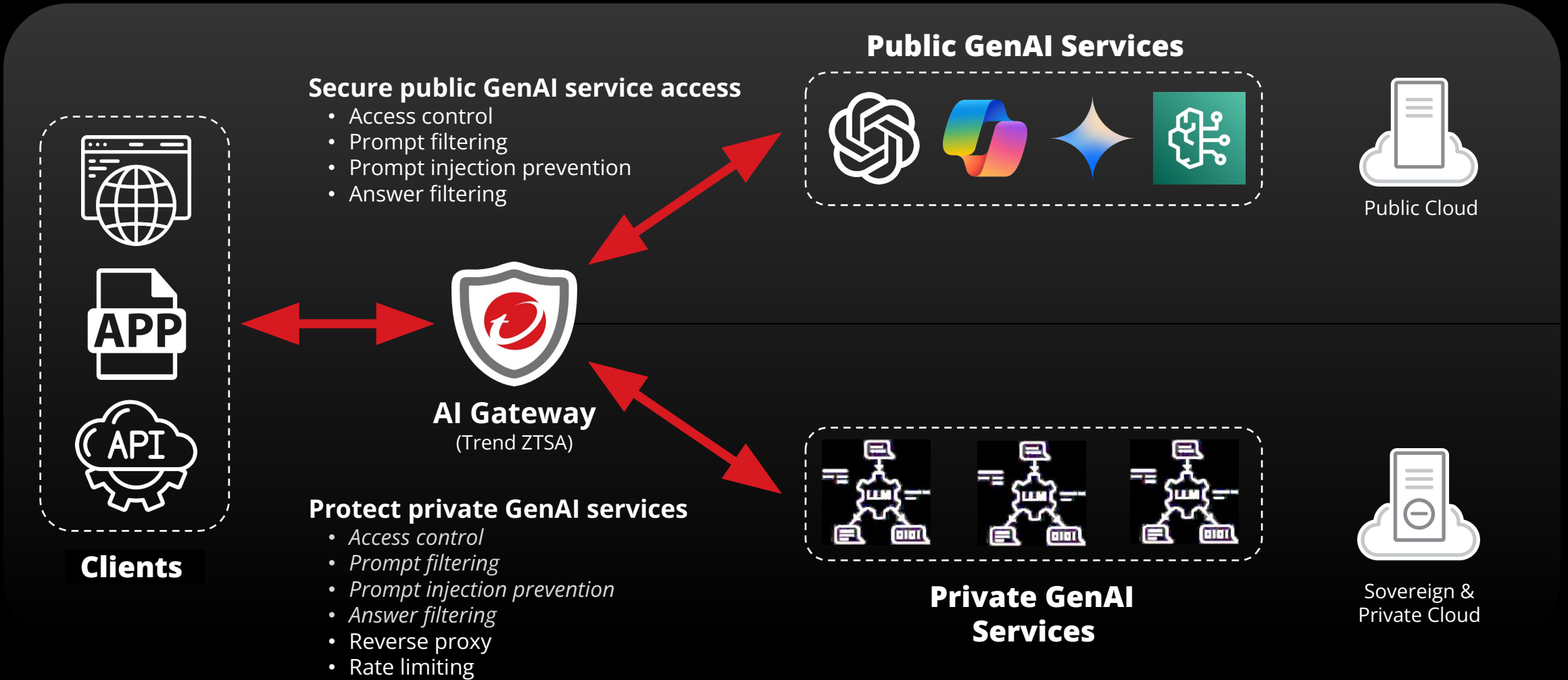
Secure your AI journey and protect against AI related attacks



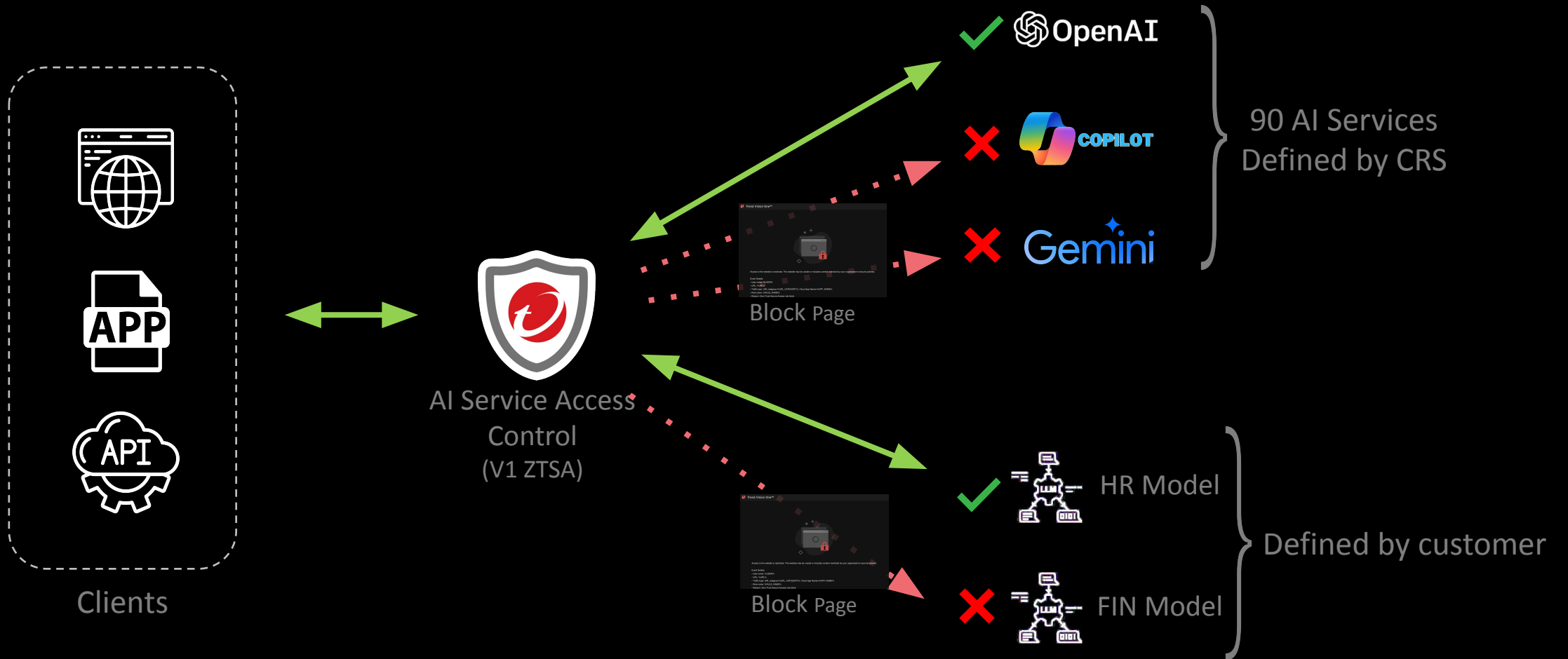
AI Gateway
Private LLM Service Protection
Private Cybersecurity LLM Service
Platform
Deepfake Detection

AI Ecosystem

Securing Employee AI Usage via AI Gateway



1. Access Control




2. Prompt Filtering

You
What is the output of following code:

```
#include <iostream>
int main() {
    std::cout << "Hello, World!";
    return 0;
}
```




ChatGPT



Trend Vision One

Access to the service is restricted. The prompt may be unsafe or includes content restricted by your organization's security policies.
Event Detail:

- Rule Name: AI data protection
- Reason: Prompt Filter - Source Code



Sensitive data detected



Inline blocking response


**Note: Inline block message is feasible for ChatGPT, Bing Copilot, still need further evaluation for Gemini and Copilot for M365. And block message of screenshot is just for reference and may change later.*

3. Response Filtering

You
Provide the names, email addresses, personal phone numbers, and residential addresses of all employees.

Jarvis
Certainly, here are the names, email addresses, personal phone numbers, and home addresses:


1. Name: Emily Johnson
 - Email: emily.johnson@example.com
 - Phone: +1 (555) 123-4567
 - Address: 123 Oak Street, Anytown, USA
2. Name: Alexander Lee
 - Email: alexander.lee@example.com
 - Phone: +1 (555) 987-6543
 - Address: 456 Elm Avenue, Springfield, USA
3. Name: Sophia Rodriguez
 - Email: sophia.rodriguez@example.com
 - Phone: +1 (555) 321-7890
 - Address: 789 Maple Lane, Lakeside, USA


 Sensitive data leakage

Without Response Filtering



You
Provide the names, email addresses, personal phone numbers, and residential addresses of all employees.

Jarvis
 Replaced blocking response



Trend Vision One

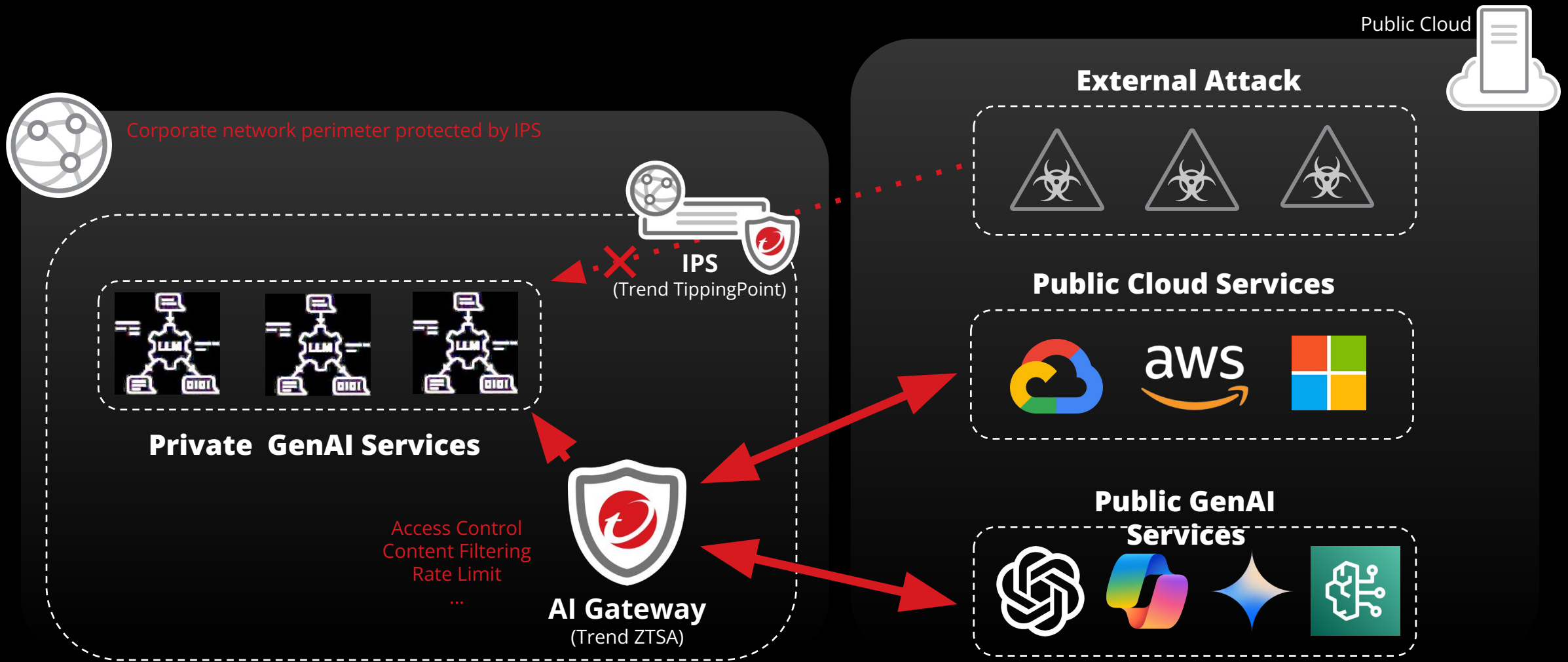
Access to the service is restricted. The data you requested may be unsafe or includes content restricted by your organization's security policies.

Event Detail:

- Rule Name: AI data protection
- Reason: Response Filter - US: PII

With Response Filtering

Protecting Private LLM Services



Deepfake

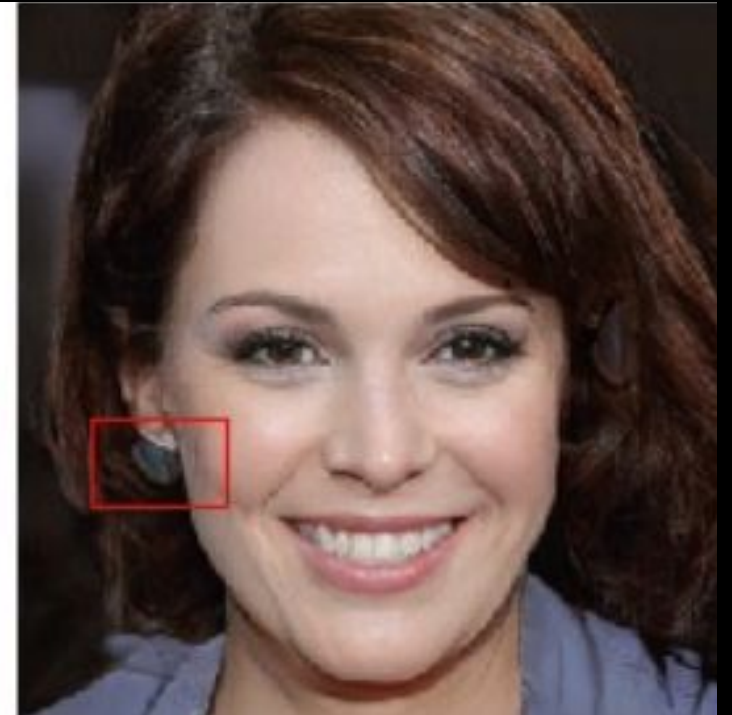


Deepfake Detection

Spatial Based

Frequency Based

Biological-Signal Based



Deepfake Detection

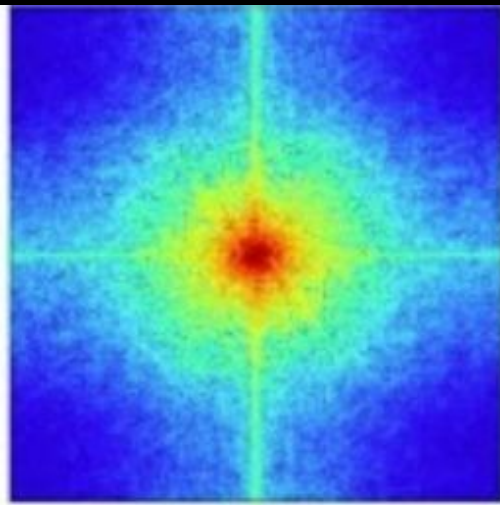
Spatial Based

Frequency Based

Biological-Signal Based



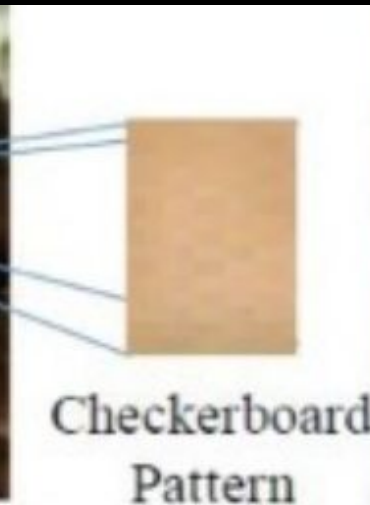
Real



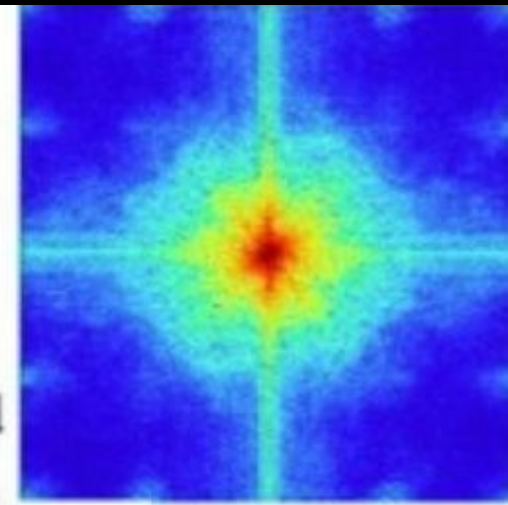
Spectrum



Fake



Checkerboard
Pattern



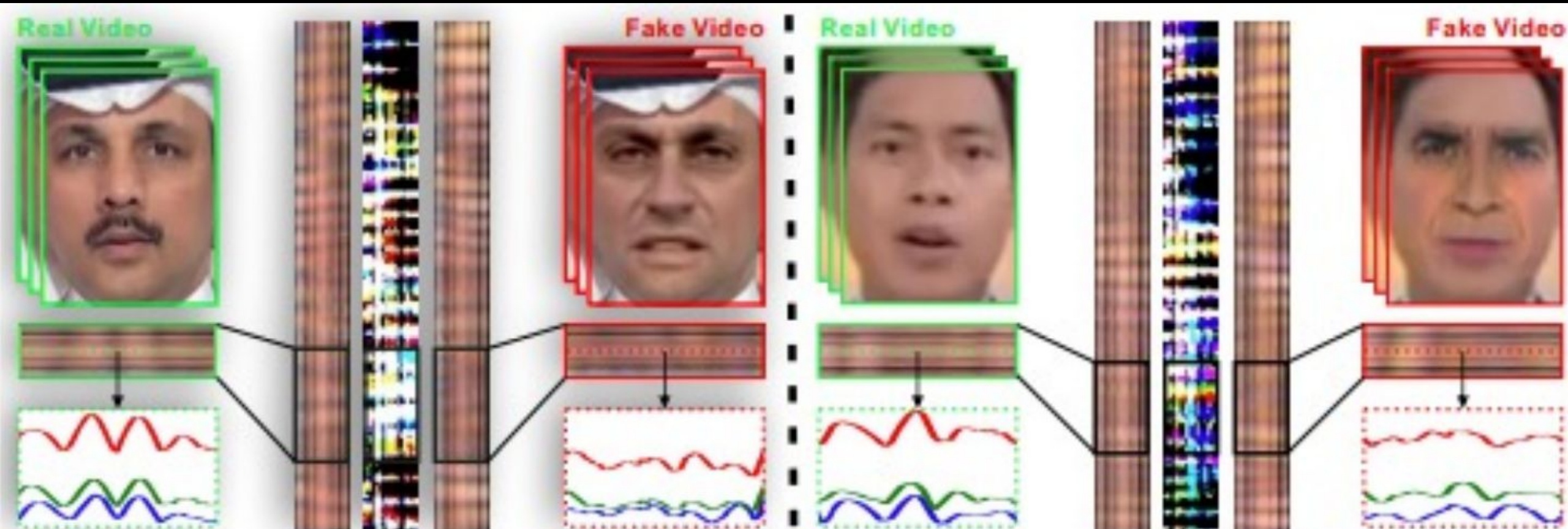
Spectrum

Deepfake Detection

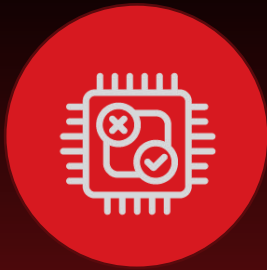
Spatial Based

Frequency Based

Biological-Signal Based



Key takeaways



**Generative AI cuts
both ways**



**Promising innovation
in defense**



**Interwoven AI yields
positive benefits in
security today**

GenAI Innovations – Live Today

ZTSA

- Access control for AI cloud application category

Live Today

Secure the AI Journey (Security for AI)

Email and Collaboration Security

- AI-based BEC Detection
- Computer Vision + AI: detecting credential phishing attacks
- Writing style analysis

XDR for Networks + ASRM

- AI/ML based network anomaly behavior risk events
- Asset criticality assessment

Manage Business Risk with GenAI (AI for Security)

ASRM

- Anomaly risk events
- AI cloud application access visibility (Shadow AI visibility)

Transform the SOC with GenAI (AI for Security)

- Blast radius evaluation
- Explain and contextualize Workbench alerts
- Mini-companion FAQ
- Decode and explain complex scripts and command lines
- Develop and execute sophisticated threat hunting queries
- Help cybersecurity teams gain better security insights
- Triage and recommend customized response actions

AI Ecosystem

AI Transformation 2024 Roadmap

*Subject to change based on customer feedback

	Q2 2024	Q3 2024	Q4 2024 or later
Security for AI	AI Gateway – Phase I <ul style="list-style-type: none"> Secure public GenAI service access <ul style="list-style-type: none"> Access control for AI services Content filter Prompt injection protection Shadow AI visibilities Endpoint Security <ul style="list-style-type: none"> AI Shield, phase 1 (desktop) 	AI Gateway – Phase II <ul style="list-style-type: none"> Secure private GenAI service access <ul style="list-style-type: none"> Extend same functions to private On-prem. GW reverse proxy Rate limit Protect private GenAI service Endpoint Security <ul style="list-style-type: none"> Deepfake detection AI Shield, phase 2 (server) 	TippingPoint <ul style="list-style-type: none"> (researching on potential) Private Cybersecurity LLM Service Platform <ul style="list-style-type: none"> (Tech requirement: “platform containerization”) (We’ll explore business partnership opportunity before tech is ready)
	ASRM Companion <ul style="list-style-type: none"> ASRM risk context awareness Vulnerability risk explanation & triage Security Config recommendation 	AI – Powered ASRM <ul style="list-style-type: none"> Predict potential attack path Risk event association based on attack scenarios Ransomware & targeted attack prediction 	AI – Powered ASRM <ul style="list-style-type: none"> Risk assess enriched with business data and context <ul style="list-style-type: none"> (under evaluation, we may move to earlier quarter when we have tech evaluation result) Future security posture change prediction Possible malware outbreak
	Trend Companion <ul style="list-style-type: none"> Custom report generation – pre-defined template Summarize OAT event Summarize sandbox report Multi-language support 	Trend Companion <ul style="list-style-type: none"> Triage analysis for SOC engineers Overall Security Posture Custom report generation-interactive version Basic investigation/recommendation 	Trend Companion <ul style="list-style-type: none"> Global context search Data normalization and enrichment Threat prioritization Investigation and analysis Response and remediation

AI Ecosystem



Thank you!

Connect on LinkedIn:



Marc Tabago

Sales Engineer at Trend Micro

