



LEADING WITH RESPONSIBLE AI

*Day 3: Risk & Reward: The
Security Perspective on
Responsible AI in
Healthcare*

How Threat Actors are Using AI Against Us

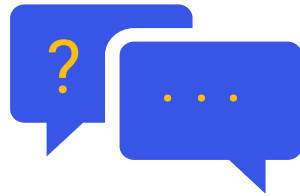
 Clearwater

Meeting Logistics



Microphones

All attendees are on mute.



Questions

Type your questions in the Q&A box.



Resources

Upcoming events, slides & resources linked.



Recording

Recording will be provided after event.



Survey

Survey will prompt at the end of webinar.

Agenda

- Welcome + Introductions
- Presentation Content: How Threat Actors Are Using AI Against Us
- Q+A



Steve Akers

Chief Technology Officer &
Corporate CISO
Clearwater



Dave Bailey, EMBA, CISSP

Vice President, Consulting
Services, Security
Clearwater

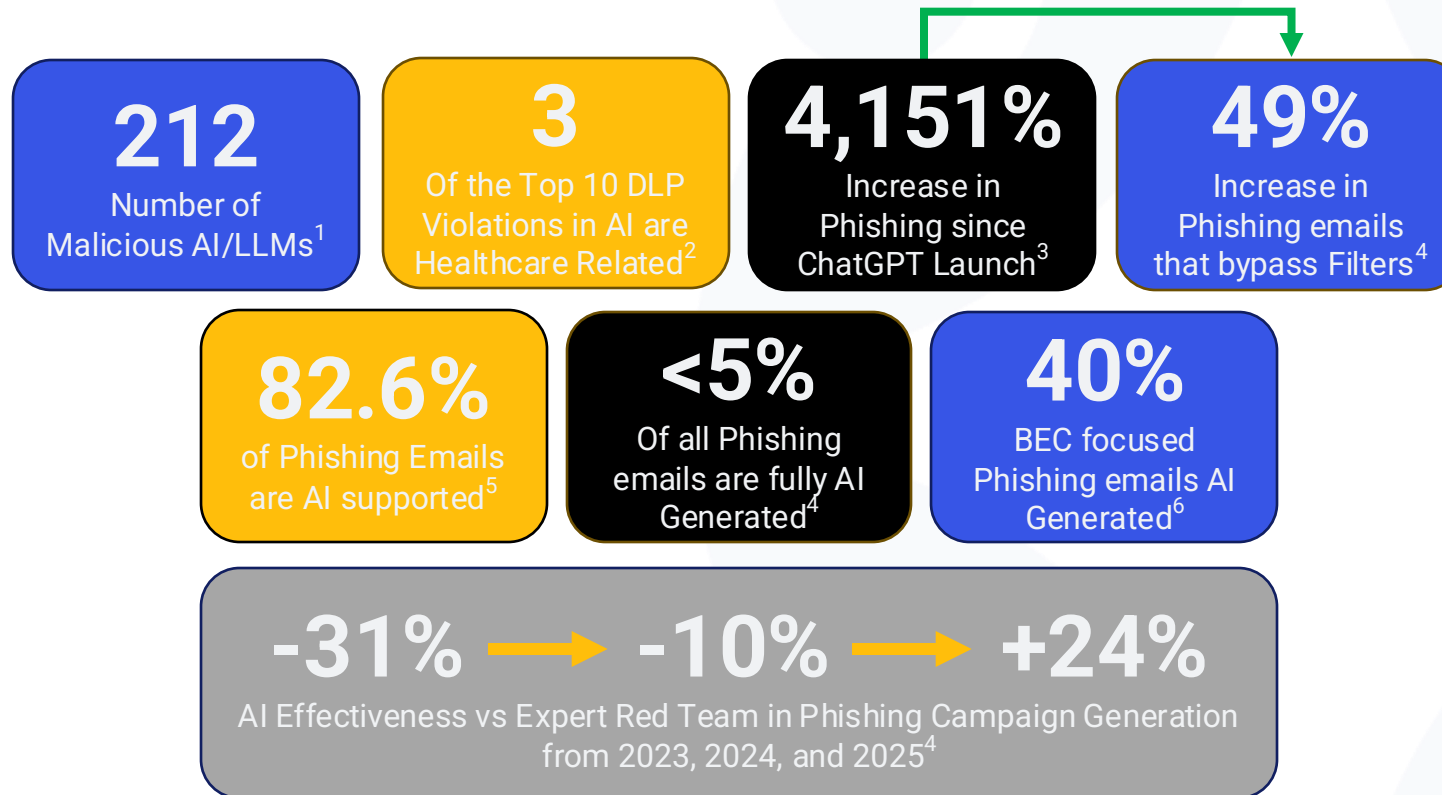
How Threat Actors Are Using AI Against Us

Steve Akers, CTO & Corporate CISO, Clearwater

Dave Bailey, VP Consulting Services, Security, Clearwater



AI in Cybersecurity



What are threat actors using AI / LLMs for?



Reconnaissance

- Target Research
- Attack Methods
- Best Path for Lateral Movement
- Vulnerabilities



Development

- Exploit Identification
- Malicious Code
- Troubleshooting
- Refinement



Automation

- Malware Development
- Mass Phishing
- Adaptive Social Engineering
- Obfuscation
- APT Lifecycle



Content Generation

- Deepfakes
- Video (Live)
- Voice (Live)
- Complete Persona



Translation and Localization

- Correct Language and Context
- Culturally aligned

Social engineering and malicious computer network operations (CNO) are primary GenAI adversary uses



“Adversaries increasingly adopted GenAI throughout 2024, particularly in support of social engineering efforts and high-tempo Information Operations campaigns”¹

How are threat actors attacking AI / LLMs?



Prompt Injection

Providing prompts designed to circumvent the rules or boundaries

Impact: Divulge information or generate questionable content



Data Poisoning

Inject fake or misleading information into training data

Impact: Accuracy or Objectivity



Evasion

Apply subtle changes to input data shared with the model

Impact: Incorrect predictions or decisions



Model Tampering

Adjust the parameters or structure of a pre-trained model

Impact: Accuracy of returned results

Two Approaches



	Jailbroken	Uncensored
Core Concept	Bypass the built-in safety restrictions or security controls of Commercial AI/LLMs	Operates without regard for harm, appropriateness, or ethical boundaries
Why this approach	Significant Technical Resources/Power Number of Parameters Access to data (PII,ePHI), Int. Property	No limits with the right training data
Output/Deliverable	Email Generation, Grammar, Translation, Leaks	Research, Content Generation, Code Analysis, Development
Delivery Method	Commercial/Free - SaaS	SaaS, Local Models
Who	ChatGPT, Gemini, DeepSeek, Copilot, Grok, Claude	WormGPT, GhostGPT, DarkBERT, Nytheon AI, 200+ More

Uncensored



- Uncensored Models are a force multiplier for Threat Actors
 - Reduced knowledge/skill requirements
 - Ability to produce output equivalent to a team of “experts”
- Easily Accessible
 - Runs on consumer grade hardware
 - SaaS Model
 - Often delivered via apps like Telegram
 - Monthly/Annual/Lifetime Subscription
 - Full Upgrades and Rapid Support
 - No Logging
 - Some even have reviews!!

Uncensored – WormGPT

CHATGPT LIMITATIONS

OPEN AI

FEATURES

- LIGHT-SPEED
- UNLIMITED CHARACTERS
- PRIVACY FOCUSED
- NO LIMITATIONS
- PERFECT CODING

PRIVACY FOCUSED

*We do not store your personal data.

WormGPT | Private. Uncensored. Exclusive.

Take WormGPT on the go with seamless performance across all platforms, ensuring maximum privacy:
Android | Linux | Windows | MacOS / iOS iPhone

Powered by cutting-edge AI, WormGPT is your go-to for hacking, coding, and elite online operations. Achieve unmatched results for any task you throw at it.

Features:

- Reasoning (Think) Feature:** Solve the toughest problems in seconds with advanced logic.
- File Upload:** Upload code or text, analyze instantly, and break all limits.
- Updated LLM Model:** Get real-time, accurate data with minimal errors.
- Next-Gen Encryption:** Keep your chats ultra-confidential with top-tier security.
- Automated Exploit Generator:** Craft custom exploits for vulnerabilities in minutes.
- Dark Web Scanner:** Scrape dark web markets for leaked data and target intel.
- Social Engineering Toolkit:** Build targeted phishing campaigns with high success rates.
- Malware Builder:** Create undetectable keyloggers, stealers, or ransomware with ease.

AND MORE!

Example Projects with WormGPT:

- Phishing Empire:** Build a fake banking site with email templates to steal credentials.

Customers can DM or Telegram us for proofs and vouches. We accept Escrow. Check our Telegram and Discord for shared proofs to ensure trust.

Payment Options:

We accept cryptocurrency for secure, anonymous transactions.

BUY NOW WORM GPT

Contact Developer / Support: <https://t.me/forsasuke>
Telegram Channel: <https://t.me/wormgptchannel> | Shut Down by Telegram Team Contact t.me/@forsasuke
Official Website: <https://wormgpt.net/>

Note:

Customers can DM or Telegram us for proofs and vouches. We accept Escrow. Check our Telegram and Discord for shared proofs to ensure trust.

[Start Contract](#)

WormGPT Pricing

To access our High Quality product "WormGPT", we offer you payment plans. More info is given below.

MOST POPULAR

\$200 | Lifetime

- Lightning Fast
- Lifetime Access
- 24/7 Support
- Works On All Devices

[Contact Us](#)

FOR DEVELOPERS

\$550 | API Monthly

- Lightning Fast
- 24/7 Support
- Works On All Devices
- Monthly Access

[Contact Us](#)

- Only Crypto Payments
- Privacy Focused
- No Limits
- Lightning Fast
- 24/7 Support
- Web API Support

[Contact Us](#)

Uncensored – Nytheon AI

- Likely Russian Backed
- Delivered via Tor or XSS
- Unifies multiple Uncensored and Jailbroken LLMs
 - Coder (Llama 3.2 18.4B)
 - GMA (Gemma 3 4.3B)
 - Vision (Llama 3 9.8B)
 - R1(RekaFlash 3 20.9B)
 - Coder R1(Qwen2 1.8B)
 - Ai (Llama 3.8B)
- Voice Enabled

NYTHEON AI MODEL SUITE

- Nytheon AI**
General-purpose
 - Advanced NLP across 100+ languages
 - Emotional tone recognition & conversation memory
 - Real-time reasoning with ultra-high accuracy
 - Up to 1 million tokens context window
 - Web Search over 50+ global search engines
 - Image generation from-text prompts (art, design, photorealism)
- Nytheon Vision**
Multimodal vision intelligence
 - Image captioning & understanding
 - OCR, object tracking, VQA
 - Scene reconstruction & design-to-code
 - AI-powered visual storytelling & animations
- Nytheon R1 Coder**
Developer's intelligent co-pilot
 - Code generation in 50+ languages
 - Autocomplete, bug fixing, refactoring
 - AI explanations + pseudocode generation
- Nytheon R1 Coder V2**
Advanced system architect & engineer
 - Full-project multi-file understanding
 - Large-scale system design & documentation
 - Integrated linting, Autotesting, CI/CD Deployment
 - Developer environment-aware and framework-smart

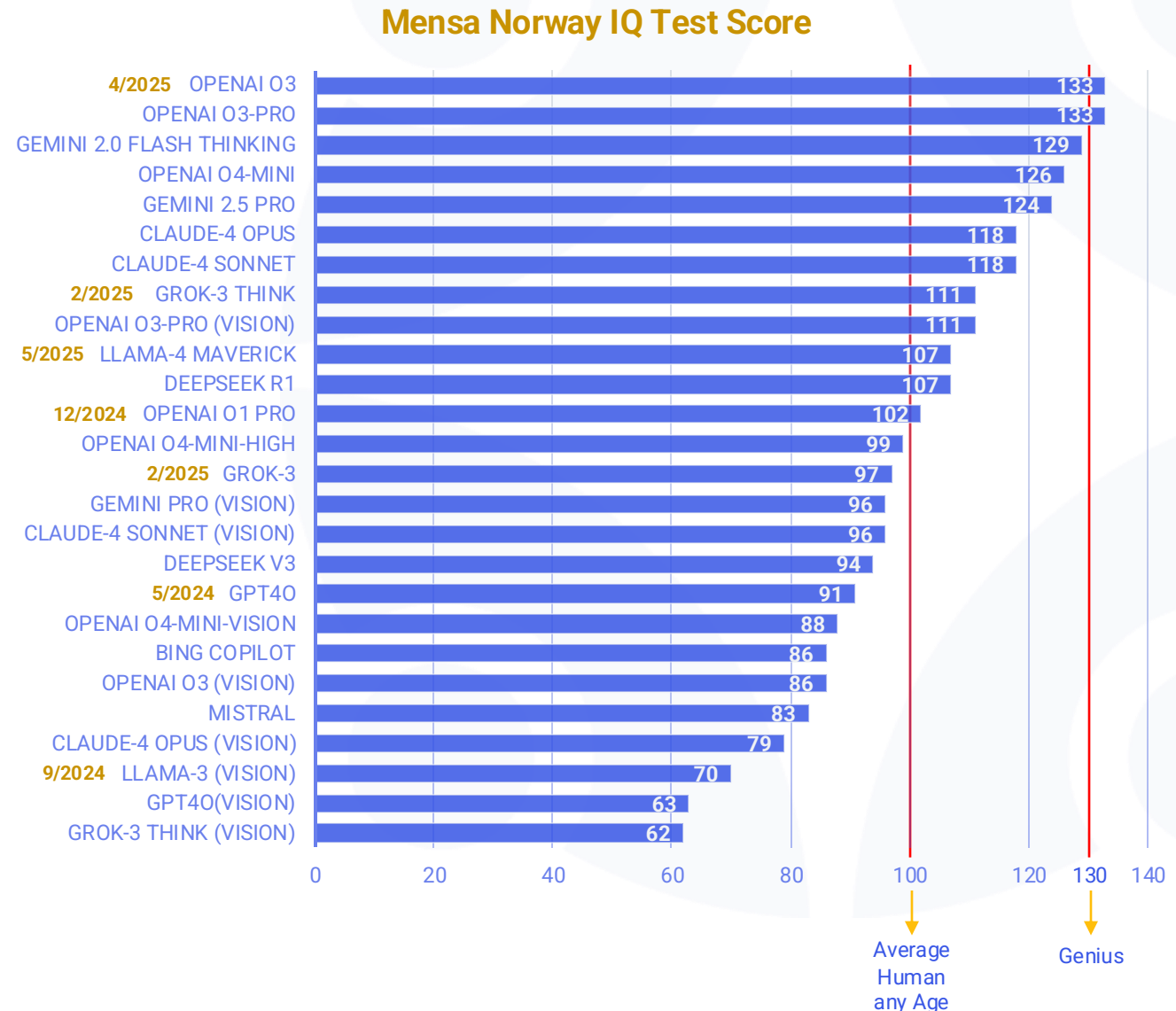
NYTHEON'S SUPERPOWERS

- Massive Context Support**
Up to 1 million tokens per interaction – perfect for books, code-bases, legal docs, research papers,
- Multimodal Mastery**
Text, image, and code – all seamlessly processed, generated, and understood.
- Hyper-Creative Image Generation**
From sketches to photorealistic renders – generate, modify and understand images with prompts.
- Self-Evolving Memory**
Learns from interaction history, adapts tone, preferences, and

```
▼ 0.6.5:
date: "2025-04-14"
▼ added:
  0:
    title: "🔑 Granular Voice Feature Permissions Per User Group"
    content: "Admins can now separately manage access to Speech-to-Text (record voice), Text-to-Speech (read aloud), and Tool Calls for each user group-giving teams tighter control over voice features and enhanced governance across roles."
    raw: "❗️🔑<strong>Granular Voice Feature Permissions Per User Group</strong>: Admins can now separately manage access to Speech-to-Text (record voice), Text-to-Speech (read aloud), and Tool Calls for each user group-giving teams tighter control over voice features and enhanced governance across roles.</li>"
  1:
    title: "🔊 Toggle Voice Activity Detection (VAD) for Whisper STT"
    content: "New environment variable lets you enable/disable VAD filtering with built-in whisper speech-to-text, giving you flexibility to optimize for different audio quality and response accuracy levels."
    raw: "❗️🔊<strong>Toggle Voice Activity Detection (VAD) for Whisper STT</strong>: New environment variable lets you enable/disable VAD filtering with built-in whisper speech-to-text, giving you flexibility to optimize for different audio quality and response accuracy levels.</li>"
  2:
    title: "📄 Copy Formatted Response Mode"
    content: "You can now enable 'Copy Formatted' in Settings > Interface to copy AI responses exactly as styled (with rich formatting, links, and structure preserved), making it faster and cleaner to paste into documents, emails, or reports."
    raw: "❗️📄<strong>Copy Formatted Response Mode</strong>: You can now enable 'Copy Formatted' in Settings > Interface to copy AI responses exactly as styled (with rich formatting, links, and structure preserved), making it faster and cleaner to paste into documents, emails, or reports.</li>"
  3:
    title: "🛡️ Backend Stability and Performance Enhancements"
    content: "General backend refactoring improves system resilience, consistency, and overall reliability-offering smoother performance across workflows whether chatting, generating media, or using external tools."
    raw: "❗️🛡️<strong>Backend Stability and Performance Enhancements</strong>: General backend refactoring improves system resilience, consistency, and overall reliability-offering smoother performance across workflows whether chatting, generating media, or using external tools.</li>"
  4:
    title: "🌐 Translation Refinements Across Multiple Languages"
    content: "Updated translations deliver smoother language localization, clearer labels, and improved international usability throughout the UI-ensuring a better experience for non-English speakers."
    raw: "❗️🌐<strong>Translation Refinements Across Multiple Languages</strong>: Updated translations deliver smoother language localization, clearer labels, and improved international usability throughout the UI-ensuring a better experience for non-English speakers.</li>"
▼ fixed:
  0:
    title: "🔒 LDAP Login Reliability Restored"
    content: "Resolved a critical issue where some LDAP setups failed due to attribute parsing-ensuring consistent, secure, and seamless user authentication across enterprise deployments."
    raw: "❗️🔒<strong>LDAP Login Reliability Restored</strong>: Resolved a critical issue where some LDAP setups failed due to attribute parsing-ensuring consistent, secure, and seamless user authentication across enterprise deployments.</li>"
  1:
    title: "🖼️ Image Generation in Temporary Chats Now Works Properly"
    content: "Fixed a bug where image outputs weren't generated during temporary chats-visual content can now be used reliably in all chat modes without interruptions."
    raw: "❗️🖼️<strong>Image Generation in Temporary Chats Now Works Properly</strong>: Fixed a bug where image outputs weren't generated during temporary chats-visual content can now be used reliably in all chat modes without interruptions.</li>"
```

Is your AI smarter than a 5th Grader?

- Most released in last 12 months
- Iterations went from below average to genius in 6 months
- AI Text Only Models far more capable than Vision Models
 - Word based reasoning



Uncensored – Rapid Gains in Capability



- Same source code was provided to an uncensored AI at three intervals
- Source code was known to have five exploitable vulns
- Same starting prompts were used
- Measuring any progress in capabilities

Response Type	18 Months	12 Months	Current
Identified Vulns	2	3	5
Suggest Exploit Options	Generic	Specific	Exact
Write Exploit Paths	No	Generic	Exact
Support exploit execution	No	No	Full Toolkit Breakdown



AI Threat Awareness



Understand Real-World Adversary Behaviors

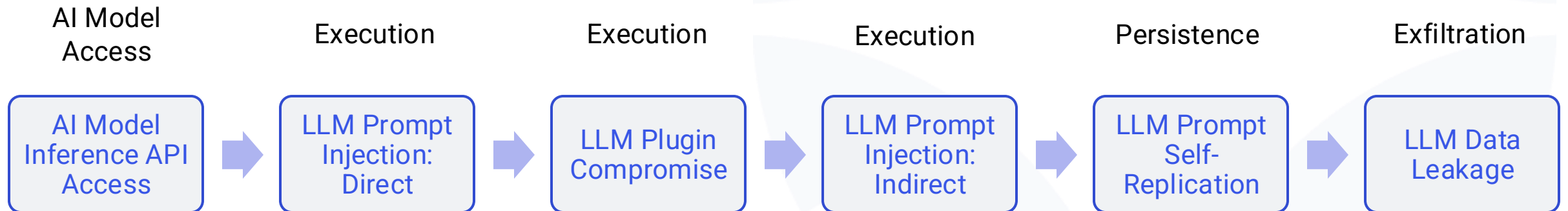
ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the [ATLAS Navigator](#).

Reconnaissance &	Resource Development &	Initial Access &	AI Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	AI Attack Staging	Command and Control &	Exfiltration &	Impact &
6 techniques	12 techniques	6 techniques	4 techniques	4 techniques	4 techniques	2 techniques	8 techniques	1 technique	7 techniques	3 techniques	4 techniques	1 technique	5 techniques	7 techniques
Search Open Technical Databases &	Acquire Public AI Artifacts	AI Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Plugin Compromise	Evade AI Model	Unsecured Credentials &	Discover AI Model Ontology	AI Artifact Collection	Create Proxy AI Model	Reverse Shell	Exfiltration via AI Inference API	Evade AI Model
Search Open AI Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	AI-Enabled Product or Service	Command and Scripting Interpreter &	Manipulate AI Model	LLM Jailbreak	LLM Jailbreak		Discover AI Model Family	Data from Information Repositories &	Manipulate AI Model		Exfiltration via Cyber Means	Denial of AI Service
Search Victim-Owned Websites &	Develop Capabilities &	Evade AI Model	Physical Environment Access	LLM Prompt Injection	LLM Prompt Self-Replication		LLM Trusted Output Components Manipulation		Discover AI Artifacts	Data from Local System &	Verify Attack		Extract LLM System Prompt	Spamming AI System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full AI Model Access	LLM Plugin Compromise	RAG Poisoning		LLM Prompt Obfuscation		Discover LLM Hallucinations		Craft Adversarial Data		LLM Data Leakage	Erode AI Model Integrity
Active Scanning &	Publish Poisoned Datasets	Phishing &					False RAG Entry Injection		Discover AI Model Outputs				LLM Response Rendering	Cost Harvesting
Gather RAG-Indexed Targets	Poison Training Data	Drive-by Compromise &					Impersonation &		Discover LLM System Information					External Harms
	Establish Accounts &						Masquerading &		Cloud Service Discovery &					Erode Dataset Integrity
	Publish Poisoned Models						Corrupt AI Model							
	Publish Hallucinated Entities													
	LLM Prompt Crafting													
	Retrieval Content Crafting													
	Stage													

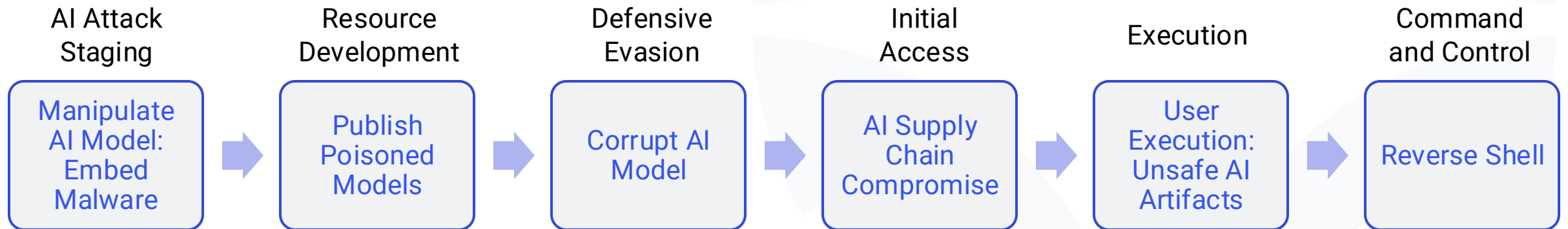
A zero-click worm designed to attack genAI

- MITRE ATLAS Case Study: The worm uses an adversarial self-replicating prompt which uses prompt injection to replicate the prompt as output and perform malicious activity. The researchers demonstrate how this worm can propagate through an email system with a RAG-based assistant.
03/05/24



Picklescan didn't flag malicious files on Hugging Face

- MITRE ATLAS Case Study: Researchers at ReversingLabs have identified malicious models containing embedded malware hosted on the Hugging Face model repository: 2025





Q&A



Contact Information



E: Steve.akers@clearwatersecurity.com

Steve Akers

Chief Technology Officer &
Corporate CISO
Clearwater



E: Dave.bailey@clearwatersecurity.com

Dave Bailey, EMBA, CISSP

Vice President, Consulting
Services, Security
Clearwater

Today's Agenda – Day 3 – Coming Next

June 25

12:00 pm – 12:45 pm CT

How Security Operations Teams Are Using AI to Fight Back



1:00 pm – 1:45 pm CT

CISO Roundtable on AI: The Good, The Bad, and The Ugly



2:00 pm – 2:45 pm CT

Ask Us Anything Breakout Session





We are here to help.

Moving healthcare organizations to a more secure, compliant, and resilient state so they can achieve their mission.



Clearwater

Healthcare – Secure, Compliant, Resilient

www.ClearwaterSecurity.com

800.704.3394

LinkedIn | [linkedin.com/company/clearwater-security-llc/](https://www.linkedin.com/company/clearwater-security-llc/)



Legal Disclaimer

Although the information provided by Clearwater Security & Compliance LLC may be helpful in informing customers and others who have an interest in data privacy and security issues, it does not constitute legal advice. This information may be based in part on current federal law and is subject to change based on changes in federal law or subsequent interpretative guidance. Where this information is based on federal law, it must be modified to reflect state law where that state law is more stringent than the federal law or other state law exceptions apply. This information is intended to be a general information resource and should not be relied upon as a substitute for competent legal advice specific to your circumstances. YOU SHOULD EVALUATE ALL INFORMATION, OPINIONS AND RECOMMENDATIONS PROVIDED BY CLEARWATER IN CONSULTATION WITH YOUR LEGAL OR OTHER ADVISOR, AS APPROPRIATE.

Copyright Notice

All materials contained within this document are protected by United States copyright law and may not be reproduced, distributed, transmitted, displayed, published, or broadcast without the prior, express written permission of Clearwater Security & Compliance LLC. You may not alter or remove any copyright or other notice from copies of this content.

*The existence of a link or organizational reference in any of the following materials should not be assumed as an endorsement by Clearwater Security & Compliance LLC.