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

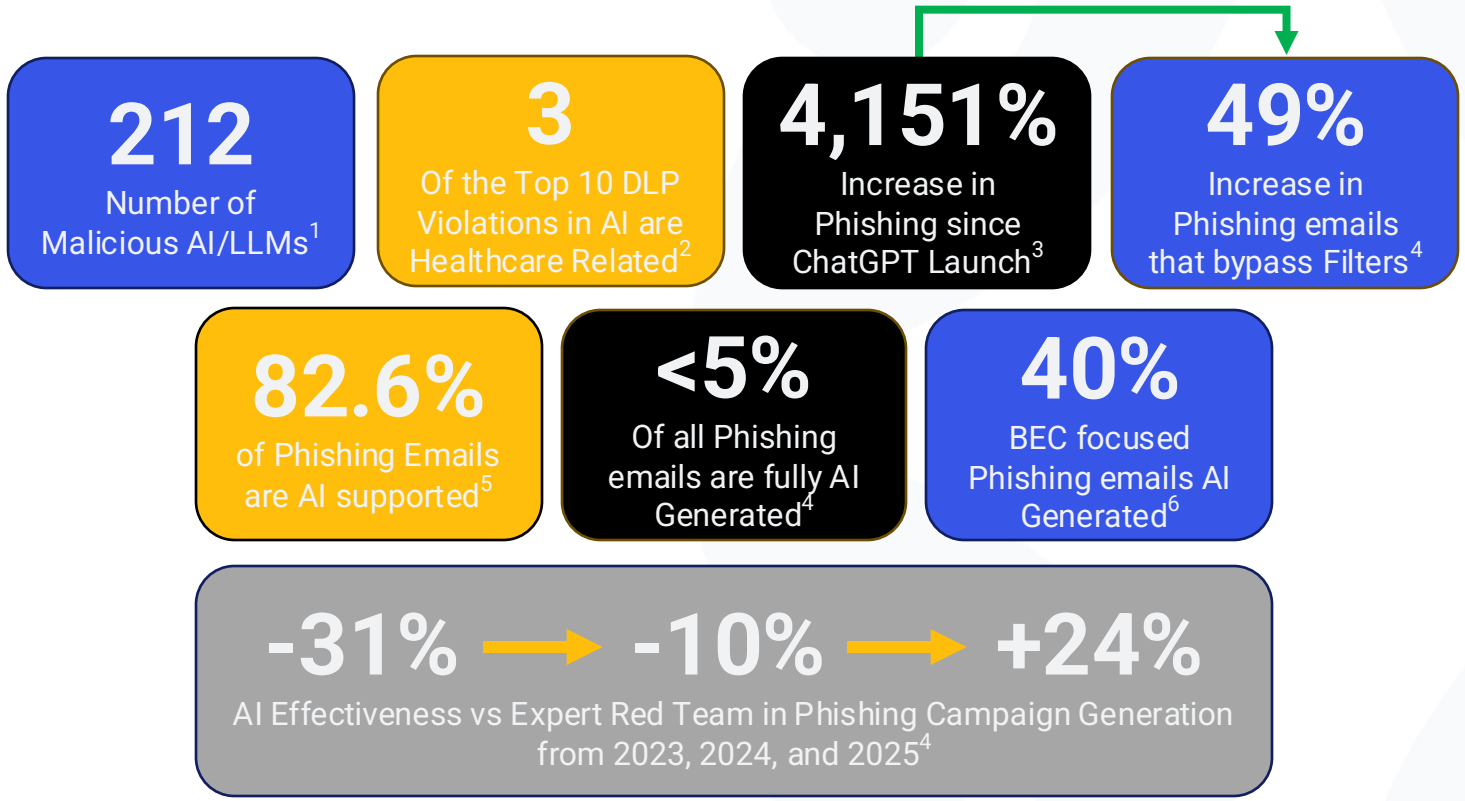
Clearwater

How Threat Actors Are Using AI Against Us

Steve Akers, CTO & Corporate CISO, Clearwater
Dave Bailey, VP Consulting Servies, Security, Clearwater

Clearwater

AI in Cybersecurity



Clearwater

[1]CybelAngel, The Dark Side of GenAI
[2]Zscaler, ThreatLabz 2025 AI Report
[3]Slashnext, Mid Year State of Fishing
[4]Hoxhunt, Phishing Trends Report
[5]KnowBe4, Phishing Threat Trends Report
[6]Vipre Security, The new face of Fraud

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
- Adaptative Social Engineering
- Obfuscation
- APT Lifecycle

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

Translation and Localization
- Correct Language and Context
- Culturally aligned

Clearwater

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]

Clearwater

1. CrowdStrikeGlobalThreatReport2025

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

Clearwater

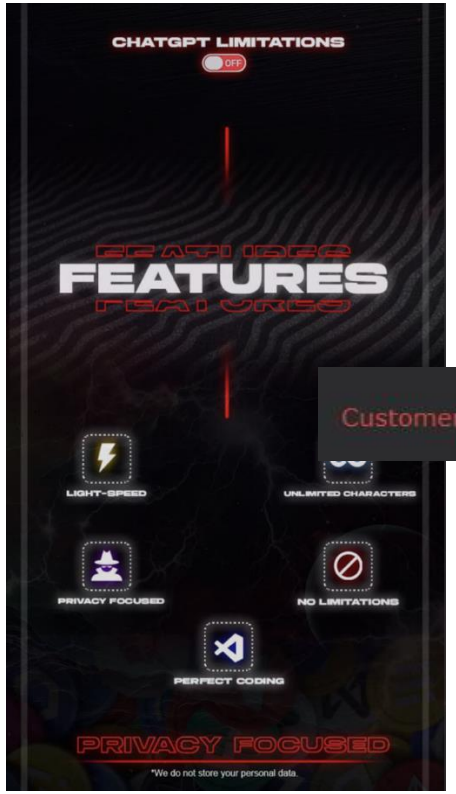
Two Approaches

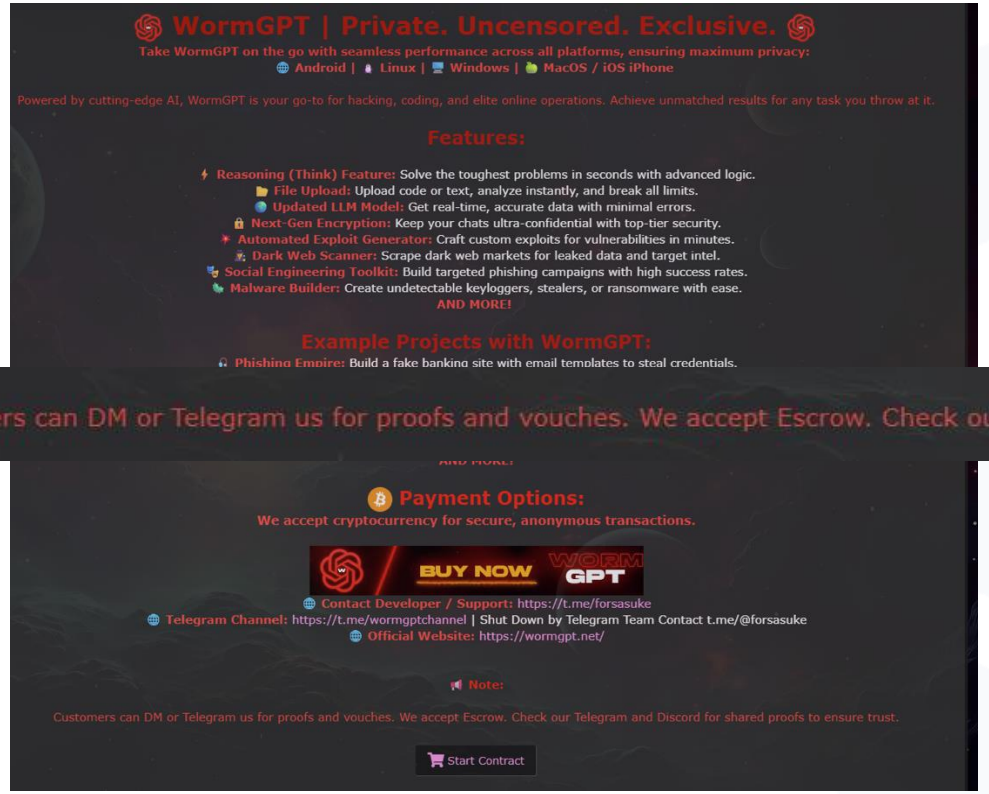
	Jailbroken	Uncensored
Core Concept	Bypass the built-in safety restrictions or security controls of Commercial AI/LLMs	Operates without regard for harm, appropriateness, or ethical boundaries
Why this approach	Significant Technical Resources/Power Number of Parameters Access to data (PII,ePHI), Int. Property	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

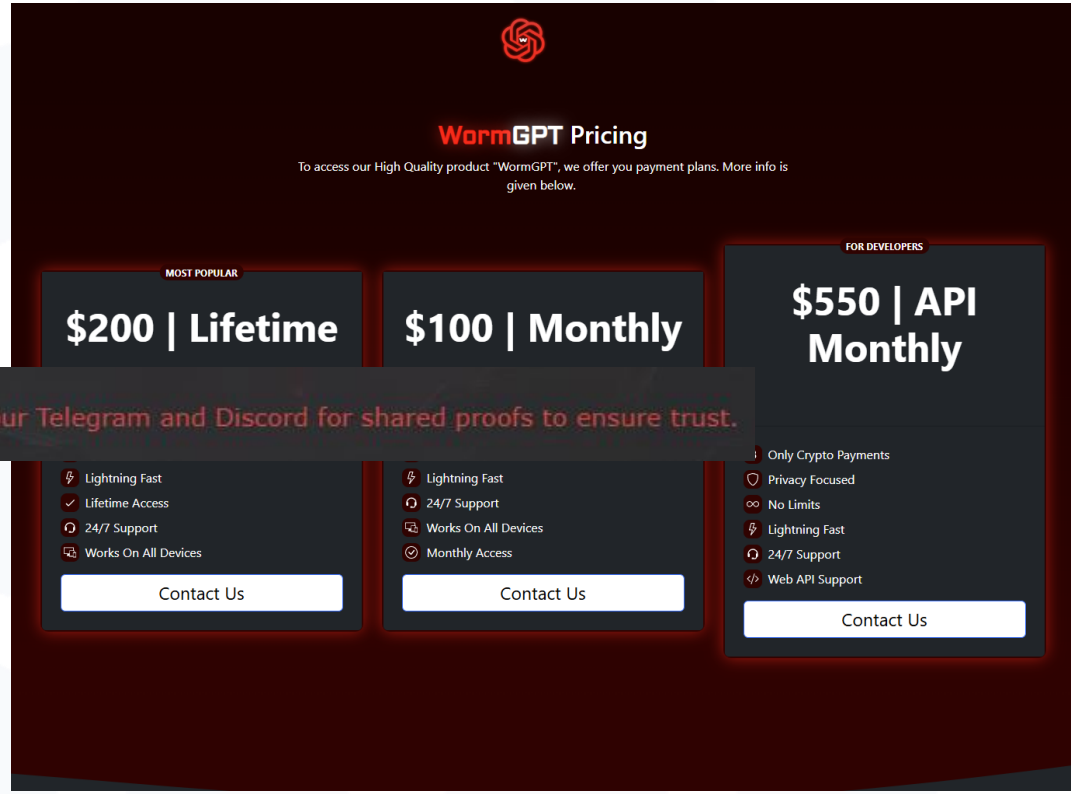
Uncensored

- 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







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

Clearwater

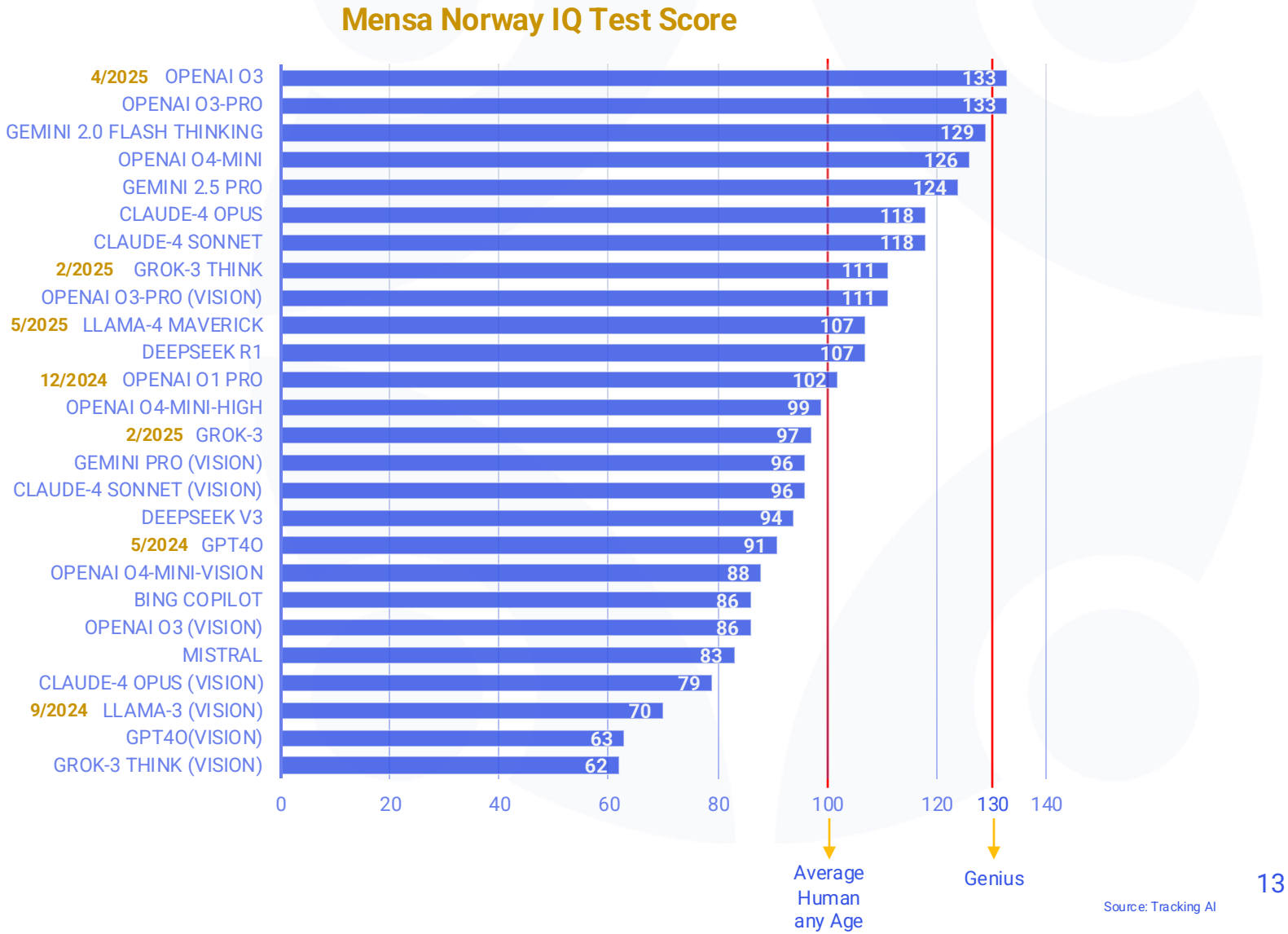
Uncensored – Nytheon AI

- 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

Clearwater

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



Source: Tracking AI

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

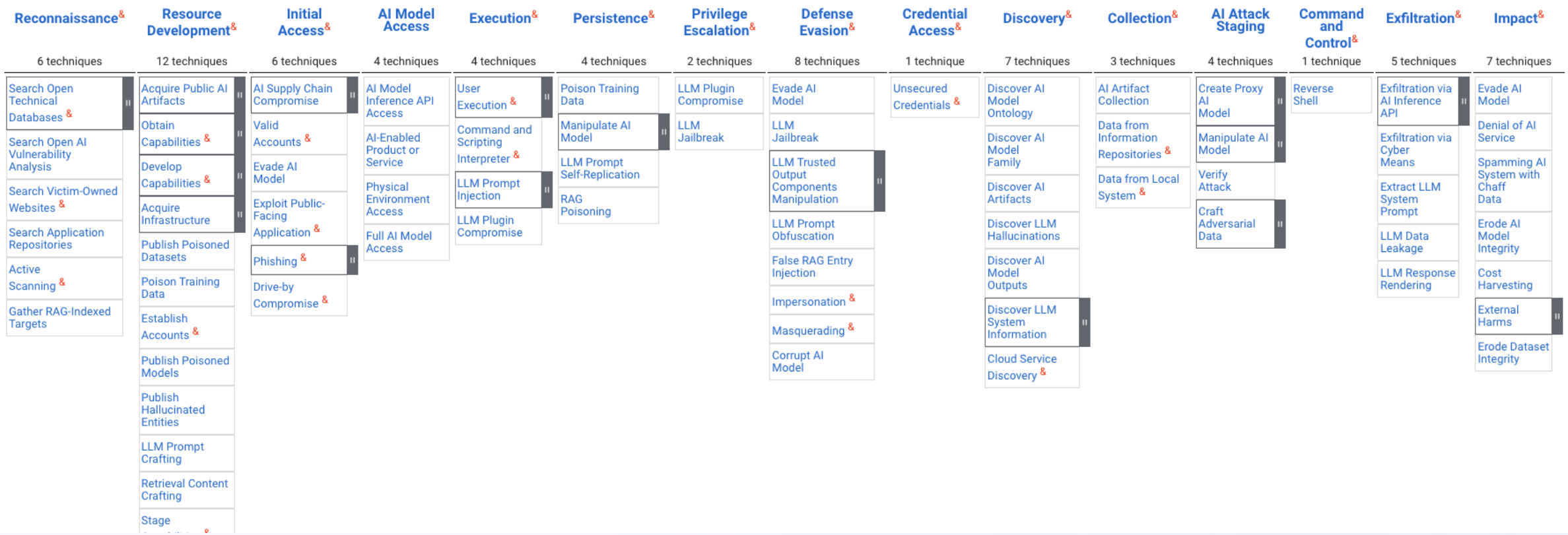
Clearwater

AI Threat Awareness

Understand Real-World Adversary Behaviors

ATLAS Matrix

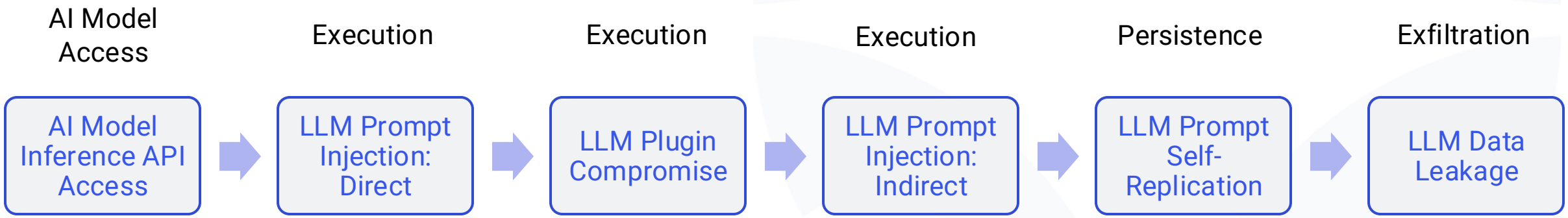
The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the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



Clearwater

A zero-click worm designed to attack genAI

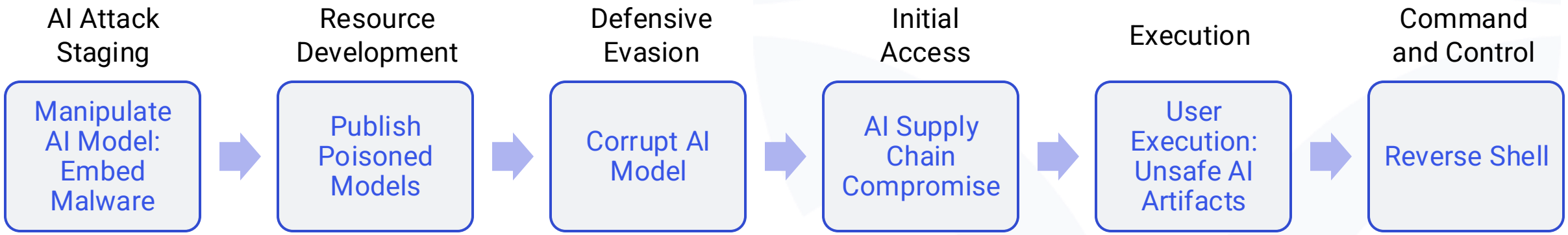
- MITRE ATLAS Case Study: The worm uses an adversarial self-replicating prompt which uses prompt injection to replicate the prompt as output and perform malicious activity. The researchers demonstrate how this worm can propagate through an email system with a RAG-based assistant. 03/05/24



Clearwater

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

Clearwater

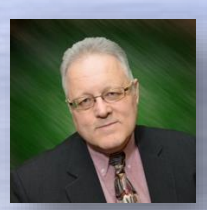
Today's Agenda – Day 3 – Coming Next

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




Clearwater

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/

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

*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.