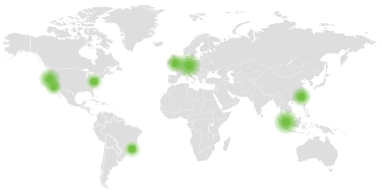
REPORT



Security Plan for "AI Threats":

LLM Attack and Defense Tests and Enterprise Response Suggestions

NSFOCUS

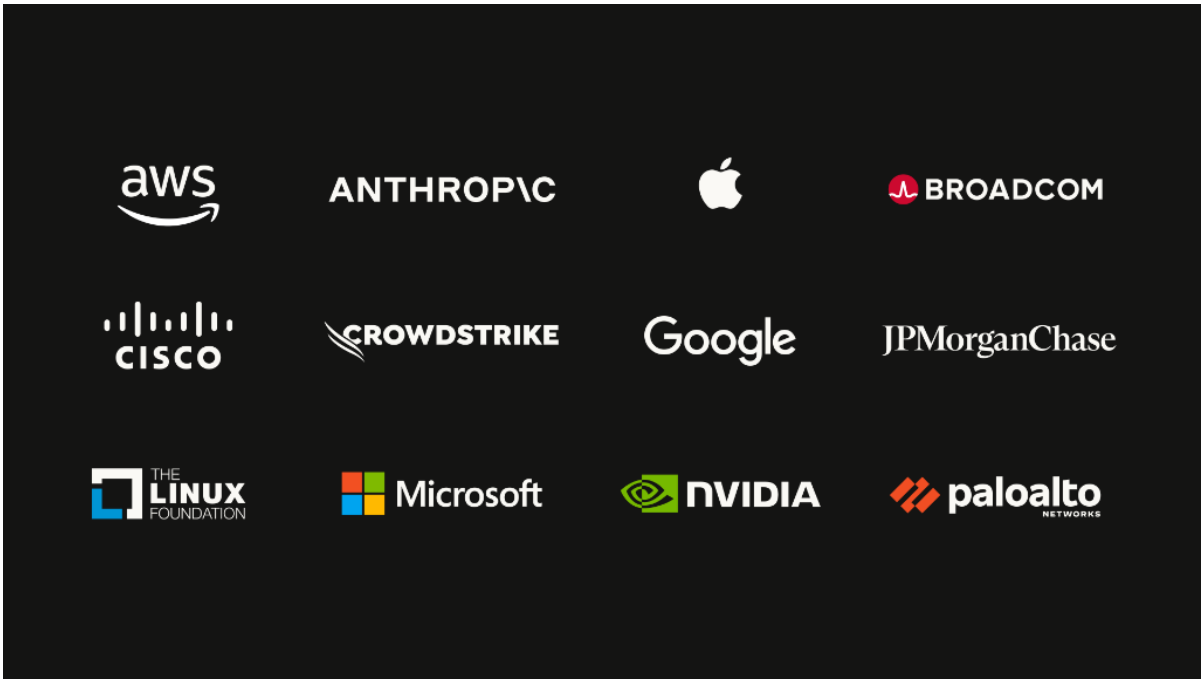
Table of Contents

Background

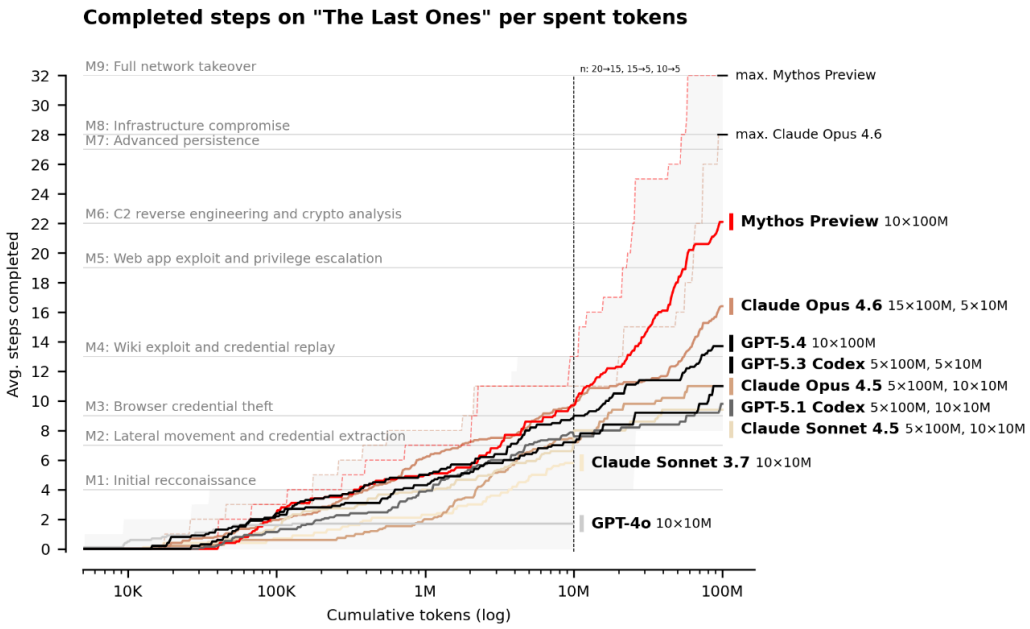
1.1 Overview of Mythos model capabilities

In February 2026, Anthropic proposed a plan to open the Claude code Review feature to provide the ability to discover and fix code vulnerabilities. In the following two months, Anthropic did not mention this again. Instead, it released its latest model on the eve of Blackhat's [un]prompted2026 conference at the end of March, which can **fully autonomously discover and exploit zero-day vulnerabilities in all major operating systems and browsers**. Including a 2027-year-old OpenBSD flaw and an FFmpeg hole that was executed 5 million times by automated fuzz testing tools but never caught.

On April 7, 2026, Anthropic officially released the AI model Claude Mythos Preview, announcing that it would not open its model capabilities to the public because the model's actual combat capabilities were too strong. At the same time, it launched a defensive cybersecurity alliance called "Project Glasswing", which **limited the use of Mythos model capabilities to only 12 founding partners and about 40 critical infrastructure maintenance organizations that had been reviewed**.



The AI Security Institute (AISI) in the UK has obtained access to the Anthropic Mythos model and evaluated the model's attack capabilities. It built The Last Ones, or TLO: a 32-step enterprise cyber attack simulation covering the entire process from initial reconnaissance to full network takeover. It is estimated that it will take 20 hours for manual completion of the simulation. The Mythos model is the first to solve the TLO problem from beginning to end, and the first to completely solve the overall cyber attack simulation problem with 32 steps. It succeeded 3 times out of 10 attempts. In all attempts, the model completed an average of 22 of the 32 steps, while the suboptimal Claude Opus 4.6 model only completed an average of 16 steps.



1.2 Release timeline of similar international models

February 5, 2026: **Claude 4.6 (Opus) was officially released**: it has powerful tool calling capabilities and can complete multi-layer intranet penetration scenarios without external guidance.

Early April 2026: **GPT-4o Cyber Policy Update (continuous grayscale)**: Strengthened interactive tools for security practitioners, focusing on single-point tasks such as disassembly analysis and protocol parsing, widening the gap with general versions.

April 7, 2026: **Claude Mythos (Preview) disclosed for the first time** that it is classified as a "strategic resource" and not open to the public due to its independent 0-Day vulnerability mining and complex attack chain construction capabilities. On April 9, Anthropic, together with Nvidia, Microsoft, Google and the Linux Foundation, officially launched a collaboration to use Mythos for global critical infrastructure defense.

April 16, 2026: **Claude 4.7 (Opus) was officially launched**: the most powerful general commercial model at present. It inherits some of Mythos' defense logic, significantly improves "honesty" and deeply integrates security audit tools. It is the first choice for industrial-grade applications.

Combined with the timeline, we can see that in April 2026, various models related to actual combat capabilities were released intensively. Among them, Mythos is irreplaceable because it is the only model granted the "autonomous vulnerability search" permission. Due to its strict control, Claude 4.7 is the "strongest" model that Anthropic has made available to the public after making Mythos' defensive capabilities into security products.

1.3 Mythos vs. opus4.6 test comparison

In terms of quantifiable indicators, Mythos shows a significant leap compared to the previous generation of publicly available models. In public data, the CyberGym indicator Mythos is 83.1% and Opus 4.6 is 66.6%; In the Firefox-related utilization experiment, Mythos successfully generated 181 usable utilizations and reached register control 29 times; In tests of about 1,000 open source repositories and about 7,000 entry points, 595 level 1 to 2 crashes were achieved, and level 5 control flow

hijacking was achieved on 10 fully patched targets. The above data shows that the improvement of Mythos's capabilities is not an occasional good result on a single task point, but a systematic increase in cross-vulnerability discovery, exploitation generation and convergence efficiency.

To facilitate horizontal comparison, the key indicators can be summarized as follows (all from public disclosure):

Indicators	Mythos	Control model	Remarks
CyberGym comprehensive indicators	83.1%	Opus 4.6: 66.6%	Official public comparison
Firefox Exploit Successful Generations	181	Opus 4.6: 2	Mythos has 29 more register controls
OSS-Fuzz class test (Level 1-2 crash)	595	Sonnet/Opus 4.6: 150-175 (interval)	Control interval from the same caliber description
OSS-Fuzz Class Test (Level 5 Control Flow Hijacking)	10	Sonnet/Opus 4.6: 0-1	Fully patched target
Coverage	**Mainstream OS + mainstream browser**	Public comparison does not give equivalent coverage results	Subject to the scope of Mythos

Core Views

1. The attacker gains an absolute asymmetric advantage. AI has greatly reduced the cost and skill threshold for discovering and exploiting vulnerabilities, making advanced cyber attack capabilities that previously required national resources within reach. The Mythos model has already shown its role in white box and real network attacks, and will become more and more stable in complex link attacks in the future. The current model could achieve zero-assistance clearance in scenarios such as CTF and cyber range, and will be closer in the future. The sub-scenario is as follows:

a) CTF: The ability of LLMs in CTF testing has shown a significant leap, with the Claude Mythos preview model achieving an accuracy rate of 73% in completing "expert" CTF tasks. Measured evidence shows that the opus4.6 model has a 79% success rate in obtaining the flag (44/56) in the challenge.
b) Range environment (black box test): The LLM has a strong sense of smell at the entrance, long-link convergence ability, information association and combined utilization ability, and can complete end-to-end compromise. Measured evidence shows that opus4.6 can be cleared in two three-layer networks at once without any auxiliary skills.
c) Code audit (white box testing): White box audit has crossed the threshold of "usable" and is approaching the critical point of "easy to use". Claude gave professional auditor-level output in minutes on both real CVEs. The actual test results show that the CVE vulnerability mining of jenkins CVE-2024-23897 (CLI arbitrary file reading) and Apache ActiveMQ CVE-2023-46604 (OpenWire deserialization RCE) were successfully obtained, and the patches are consistent with the official ones.
d) Real network confrontation: Considering that national-level confrontation emphasizes both concealment and persistence, the current model has not yet been implemented, so its capabilities need to be verified in a strong confrontation environment. However, when applied in national-level confrontations, its information collection and analysis capabilities, including phishing forgery, have been fully verified, and the depth and breadth of its information collection will exceed traditional manual capabilities.

2. In the field of offensive and defensive confrontation, its confrontation results are jointly determined by model capabilities + tools and agents, and there is a superposition relationship. Model capabilities determine the foundation of capabilities in the offensive and defensive confrontation fields, but the application framework built around the model, including Agent, prompt, Skills, tool sets, etc., will also significantly affect the release effect of model capabilities. The stronger the correlation ability, the easier it is for the gap between models to be narrowed. The current AI may perform well in normalized penetration testing, but it does not mean that it is ready to conduct penetration testing under advanced "maze" conditions such as trapping and honeynets. Therefore, when the model correlation is determined, from the perspective of raising the upper limit of offensive and defensive capabilities, we can focus on investing in relevant capability reserves and actively carry out the development and deployment of adversarial intelligent agents.

a) Context management: Because real network attacks are multi-stage and long-term tasks, the model needs to remember previous discoveries, avoid repeated operations, and maintain the coherence of attack paths.
b) Gradually accumulate high-quality skills to form knowledge accumulation: Under the same model, combining Skill with Agent will significantly improve performance in specific tasks, but the essence of Skill is closer to "ability patching" than "ability replacement".
c) Multi-model routing: There are many subdivision scenarios in the attack and defense fields, and the model advantages are also different. Therefore, it is necessary to achieve scale and cost optimization based on multi-model collaboration and routing.

3. The Mythos model will play an important role in various fields of security, as shown below:

a) Transformation of daily security operations from "passive defense" to "active hunting"
   - Active vulnerability clearing: Actively identify the exposure surface and attack path of daily operations, actively discover and automatically reproduce vulnerabilities, and realize "vulnerability discovery and exploitation".
   - Incident response: Realize real-time deduction of the risk radius of security incidents, and use asset and vulnerability information to countermeasures against assets that attackers may endanger.
   - Cost optimization: Combined with the business launch, real-time online testing that cannot be achieved by traditional manual online evaluation is realized.

b) CI/CD, from "fix after the fact" to "development as defense"
   - Realize development as defense through deep semantic scanning + self-verification and repair: understand code intent through deep semantic scanning, track vulnerabilities across files, combine multi-model collaboration to improve detection accuracy, identify complex business logic vulnerabilities, independently generate verification cases and fix patches, adapt pipeline rapid iteration, and realize security left shift.
   - Pipeline full node protection: covers the entire CI/CD process nodes, from code submission, construction, testing to deployment, embeds security detection in real time, automatically intercepts vulnerable codes and non-compliant configurations, and prevents vulnerabilities from flowing into the production environment; at the same time, generates detection reports and optimization suggestions, and links the development team to approve automated security fix patches to achieve closed-loop repair.

4. The Mythos model simulates the entire process of human security experts from code review to vulnerability exploitation and full-chain network penetration, and is executed at machine speed and scale and presents the dual characteristics of exponential compression and scenario differentiation:

a) Comprehensive vulnerability mining with tens of millions of lines of code used to take security experts months, but Mythos costs less than $20,000 and can be completed in a few hours. Single vulnerability identification can be completed within minutes and the cost is less than $50; AI discovers and verifies vulnerabilities at machine speed, and the number of subsequent vulnerability disclosures may far exceed historical trends;
b) AI compresses the weaponization time of vulnerabilities to hours, while patch cycles are still measured in days/weeks, putting defenders at a systemic disadvantage;
c) Mythos can autonomously integrate 3-5 vulnerabilities to form a multi-stage attack sequence. After obtaining the target and network access rights, complex multi-stage attacks can be completed without continuous human intervention; from initial access to domain administrator privileges, AI can complete it within hours, while human penetration testers usually need days.

5. Open-source supply chains are emerging in batches on the 0-day, and the timeliness of intelligence and monitoring responses will be particularly critical in the future. Compared with the entities participating in the Glass Wing program, non-participating companies and countries face more severe zero-day vulnerability challenges. The core focuses on:

a) Mythos' asymmetric advantages in the number and speed of zero-day mining and vulnerability exploitation will lead to zero-day vulnerabilities in infrastructure such as Linux and Docker. Due to the lag in obtaining vulnerability intelligence, it is impossible to share planned vulnerability warning and disposal resources, and the defender will be trapped in a passive situation.
b) The risk of critical infrastructure exposure is prominent. Glass Wing plans to report a large number of vulnerabilities in such facilities, which may be exploited by those with ulterior motives to launch precise attacks against non-participants, and the speed of vulnerability exploitation far exceeds the ability of manual disposal.

Response Suggestions

Large language models represented by Claude Mythos have greatly reduced the cost and skill threshold for discovering and exploiting vulnerabilities, making advanced cyber attack capabilities that previously required national resources within reach. At the same time, it also compresses vulnerability discovery from weeks of human experts to hours, and the time window for vulnerability discovery and exploitation is sharply narrowed. The core pain points of network security have undergone a structural shift: The time window from vulnerability disclosure to weaponization (generating attack payload) is infinitely compressed. The introduction of "AI versus AI" competes for the defense time window, fundamentally narrowing the fatal gap between "vulnerability discovery speed" and "organizational response speed".

3.1 Attack-defense confrontation

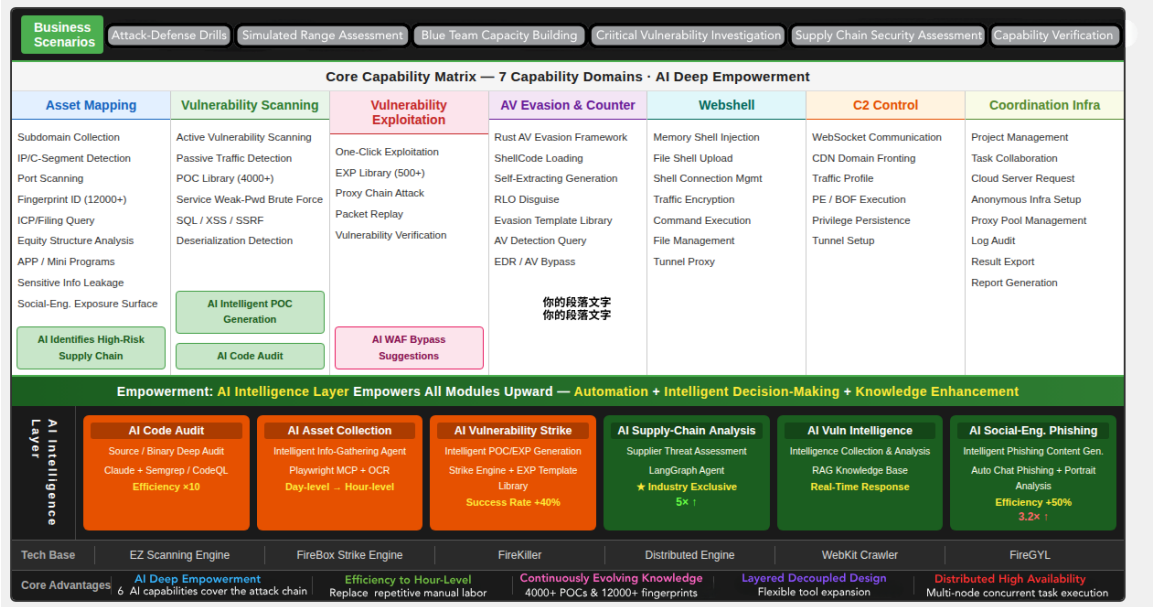
In the field of confrontation between large-scale cyber exercises and APT, AI is currently better at non-confrontational scenarios such as information collection and analysis, white box vulnerability mining, etc. It is easier to be caught in a strong contested environment with large "movement" in penetration combat. At present, it is mainly in an auxiliary attack role. The defender can introduce the perspective of AI attackers to assist in defense.

(1) Sort out the list of important suppliers and strengthen supply chain security control

Select key core and important supply chain units from the massive supply chain system for reminders, strengthen supplier control by tightening minimum permissions and implementing a zero-trust micro-isolation architecture, thereby minimizing the risk of losing points due to supply chain boundary breakthroughs and horizontal movement paths in the intranet during actual combat exercises.

(2) Build an AI+Blue Army integrated platform to achieve normalized attack and defense drills

In the face of open source supply chain risks and asset exposure risks, we actively sort out the risks of key open source components, focus on critical infrastructure such as Linux, Docker, open source databases, third-party software and open source components, and actively use advanced LLMs for vulnerability mining and real network attack and defense confrontation to improve network defense resilience.



AI empowers actual combat attack and defense, building an intelligent platform from information collection to boundary breakthrough, and then to vulnerability crackdown

3.2 Development and Security Operations

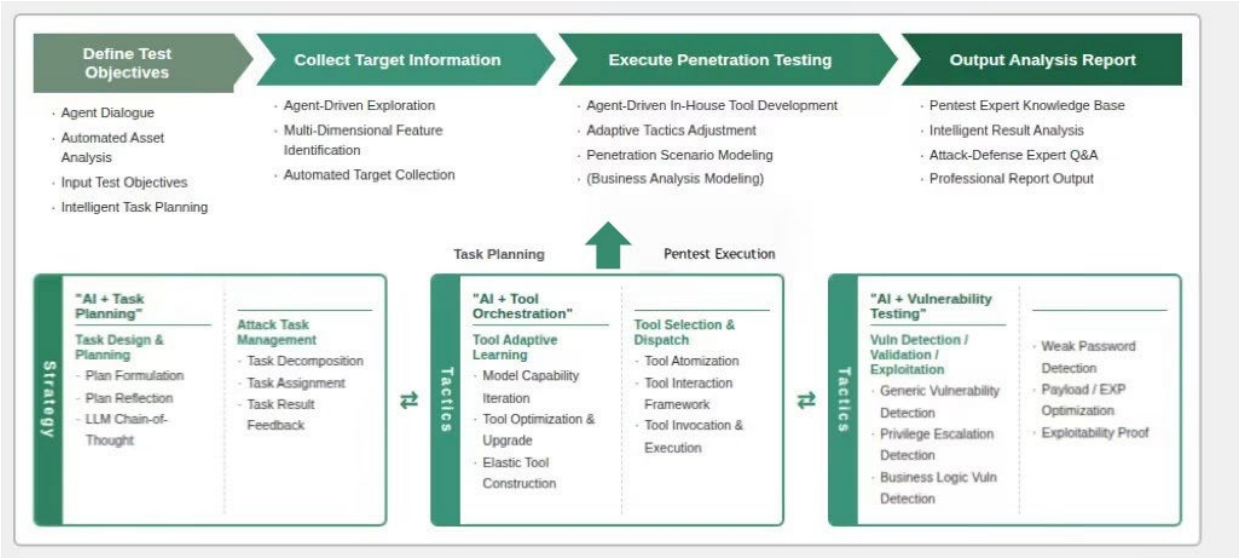
In the development process, AI capabilities are integrated into the R&D pipeline, using AI's advanced programming capabilities to strengthen the implementation of secure coding specifications in the coding stage, and using AI's vulnerability mining capabilities to fully test and reduce vulnerabilities at various stages before the system goes online. In daily operations, an AI-driven security operation system is built to shorten the vulnerability response time window.

(1) Introduce domestic advanced AI security agents to carry out continuous and active vulnerability mining of core systems

For important information systems, deploy domestic advanced AI security agents to carry out continuous vulnerability mining, discover their own vulnerabilities before attackers, and seize the initiative.

(2) Hyper shift-left security, automatic interception in the ci/cd stage

Considering the changes in the future AI coding work paradigm, introduce AI coding specifications to reduce the probability of introducing vulnerabilities due to human negligence. The development team should integrate the AI security testing agent directly into GitHub Actions or GitLab CI/CD pipelines. Through AI-enabled code auditing and penetration testing, if serious security flaws (such as SQL injection or unauthorized access) are found, they will automatically mark the PR as "need to be fixed" and prevent virus-carrying code from entering the main branch during the merge phase.



AI Attack and Defense Agents Empower Full-Process Penetration Testing

(3) Build software and AI bill of materials (SBOM/AIBOM)

Comprehensively sort out and establish a software and AI bill of materials (SBOM/AIBOM), force the introduction of SBOM and AIBOM in the development flow and procurement life cycle, accurately grasp all third-party microservices, containers, open source code libraries, AI models and their training data sources integrated within the system, and instantly locate the affected business scope when the underlying open source components explode zero-day vulnerabilities discovered by models such as Mythos. Timely response and disposal.

(4) Strengthen the defense-in-depth system and strengthen the confrontation capabilities of borders and terminals

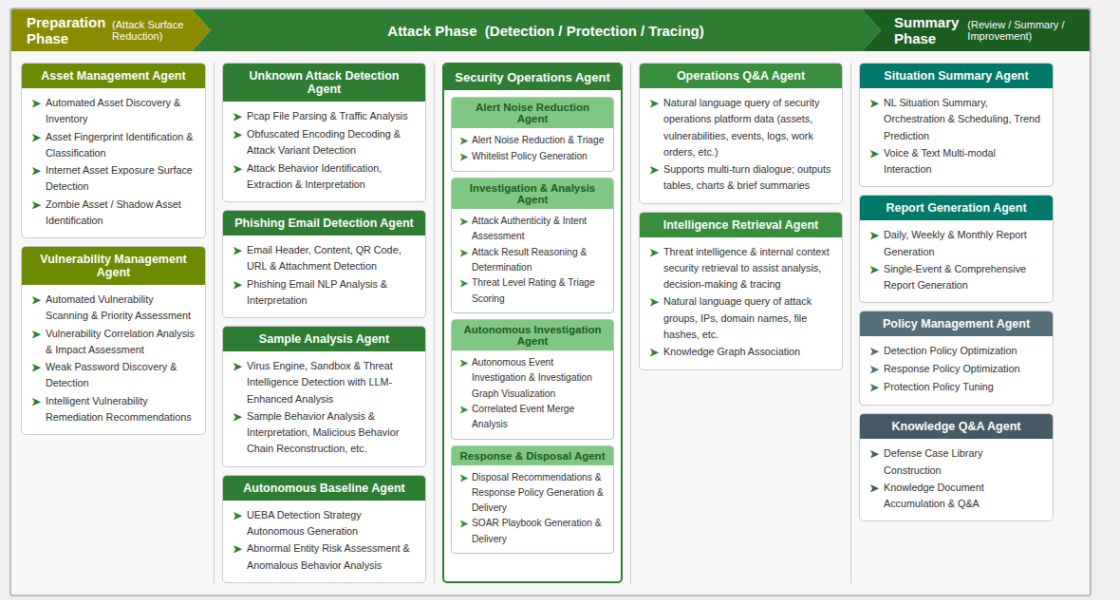
Given AI's intuition or sense of smell for human experts in offensive and defensive confrontations, it still needs to make attack attempts in a "well-defended system" to trigger security alarms. The defense-in-depth mechanism is still one of the most effective and core strategic means to fight against advanced artificial intelligence network threats represented by

Mythos. The network side strengthens network isolation, zoning and domain division, as well as strict monitoring mechanisms such as honeypots, abnormal traffic analysis, monitoring audits, etc., to build a defense-in-depth mechanism. Similarly, on the assumption that system vulnerabilities will inevitably be quickly discovered by AI, the focus of defense should shift to controlling the scope of impact after the vulnerability is exploited, returning to the construction of security infrastructure, implementing strong access control and authentication, fine-grained network isolation (micro-segmentation) and comprehensive logging to strengthen the terminal side's confrontation capabilities. Even if AI discovers a zero-day vulnerability in an edge service, Perfect network isolation can also delay or prevent the time it takes to build a multi-step attack chain across hosts and network segments.

(5) Build an AI + security operation system to strengthen threat monitoring and response mechanisms

Due to the asymmetry between attack and defense, it is still necessary to be prepared for vulnerabilities. Focusing on core important assets, automated triage and judgment based on business context, small models are responsible for preliminary filtering and high-frequency noise reduction of massive heterogeneous logs at the billion level per day, while LLMs are responsible for in-depth reasoning and tracing of complex attack chains, building an AI+Security Operations (XDR) platform, and comprehensively recording important operational behaviors. Unified collection of multi-source data (logs, traffic, terminal behavior), using AI to identify attack behaviors scattered across different nodes, improving the accuracy and speed of attack detection.

Based on AI Agent-driven automatic response, in the face of highly repeated and clearly structured attacks, authorize AI agents to perform automatic containment (such as isolating affected network segments, banning abnormal IP or service accounts), achieve closed-loop blocking, and form a machine-level response closed loop.



Multi-agent drives safe operation: a panoramic view of AI empowerment from preparation to summary

(6) Industry collaborative intelligence early warning to form an industry joint prevention and control mechanism

With the help of intelligence early warning agent capabilities, it can capture intelligence from dark web and security forums, integrate information, and automatically extract samples to form an industry threat intelligence sharing platform. The vulnerabilities mined by AI in various institutions are shared in real time (de-identified), eliminating the time difference of single-point defense, so that attackers cannot use the "information gap" to break them one by one.

(7) Strengthen the R&D of supporting agents for LLMs based on business context

AI LLMs and suitable agents can play a better role in combination. Domestic LLM manufacturers mainly focus on improving the capabilities of the models themselves, and agent supporting products are relatively lagging behind. It is recommended that relevant departments guide domestic manufacturers to build an agent tool ecosystem that adapts to themselves while improving model performance in relevant policies.