CSAI Foundation

Cloud Security Alliance AI Safety Initiative

Automated Exploit Generation: LLMs Cross the Threshold

From Vulnerability Discovery to Weaponization – Understanding the New Threat Landscape

Unofficial AI-assisted Research

2026-04-02

Table of Contents

Executive Summary

For years, practitioners debated whether large language models (LLMs) would materially change the offense-defense balance in cybersecurity — or whether their outputs would remain useful only as coding assistants and documentation tools. The evidence accumulated across 2024 and 2025 makes a compelling case that this debate has largely been answered. LLMs can now autonomously discover exploitable vulnerabilities in production software, generate working exploit code from vulnerability advisories, execute multi-stage network attacks across realistic environments, and outperform most human teams in competitive hacking exercises. The threshold from vulnerability discovery to weaponization has not merely been approached — it has, by CSA's assessment of the available evidence, been crossed.

This whitepaper examines the research trajectory that brought us to this point, documents the real-world incidents that confirm theoretical capabilities are being actualized, and assesses what this shift means for enterprise defenders, security researchers, AI providers, and regulators. A central finding is significant: the cost and skill required to move from knowledge of a vulnerability to working exploit code has declined substantially. What previously required a skilled security engineer spending days or weeks — understanding a vulnerability class, developing a proof of concept, chaining preconditions — can now be initiated by a motivated attacker with access to a capable LLM, a CVE description, and API costs measured in tens of dollars, as demonstrated in the Incalmo research reviewed below.

The implications extend beyond individual organizations. When automated exploit generation becomes broadly accessible, the economics of vulnerability disclosure change. The attack surface of AI systems themselves expands. The gap between patch release and mass exploitation narrows. And the threshold for what constitutes a capable threat actor has expanded considerably.

This document provides security teams and executives with a grounded picture of where capabilities stand today, how they are likely to evolve, and what defensive and governance responses are necessary.

Introduction: The Shifting Landscape of Offensive AI

The history of automated vulnerability research is long, predating large language models by decades. Fuzzing, symbolic execution, and constraint solvers have automated aspects of bug discovery since the 1990s. What distinguishes the current generation of LLM-based tools is not automation per se — it is the breadth and generality of the capability. Prior automated systems were narrow specialists: a fuzzer could

find memory corruption in C programs, but could not reason about business logic, interpret vulnerability advisories in natural language, or chain together a multi-stage attack across heterogeneous systems. LLMs, by contrast, bring general-purpose reasoning to the problem.

This generality is consequential for two reasons. First, it lowers the expertise required to deploy offensive tools. A threat actor who previously needed deep knowledge of memory corruption primitives or web application security can now augment their capability substantially with natural language prompting. Second, it compresses the timeline from vulnerability publication to exploitation. In a world where exploit development required specialist skill and days or weeks of work, organizations had a defensible window after patch release. That window is shrinking.

The progression from research curiosity to operational concern has been rapid. In April 2024, researchers at the University of Illinois Urbana-Champaign published a widely cited study demonstrating that GPT-4 could autonomously exploit 87 percent of a set of real one-day vulnerabilities when provided with CVE descriptions [1]. By mid-2025, autonomous red teaming frameworks were executing multi-stage attacks across multi-host enterprise-like environments in under an hour, at costs below fifteen dollars in API credits [2]. In August 2025, DARPA concluded the two-year AI Cyber Challenge (AIxCC), in which competing teams' autonomous systems identified 86 percent of synthetic vulnerabilities embedded in millions of lines of real code — and discovered 18 previously unknown real-world flaws in the process [3]. By late 2025, incidents surfaced in which real attackers were using commercial LLMs to generate functional exploit code against government targets [4].

This whitepaper traces that arc: from benchmarks and research demonstrations to operational capability and real adversarial use. It does not predict when fully autonomous offensive AI will dominate the threat landscape. It documents that meaningful elements of that capability already exist, and that the trajectory points toward broader accessibility and more capable systems.

The Research Baseline: Benchmarking LLM Exploit Capabilities

The UIUC One-Day Exploit Study

The most cited quantitative inflection point in the LLM exploit literature is the April 2024 paper "LLM Agents can Autonomously Exploit One-day Vulnerabilities" from the University of Illinois Urbana-Champaign [1]. The authors — Richard Fang, Rohan Bindu, Akul Gupta, and Daniel Kang — constructed a

dataset of 15 real one-day vulnerabilities spanning web application flaws, container vulnerabilities, and vulnerable Python packages, with more than half rated as high or critical severity. Eleven of the 15 arose after GPT-4's training cutoff, ruling out memorization as an explanatory factor.

Given only the CVE advisory text, GPT-4 successfully exploited 87 percent of the vulnerabilities — a striking result given that no other model tested, including GPT-3.5, open-source LLMs, and established scanning tools such as OWASP ZAP and Metasploit, succeeded on any of the vulnerabilities. The contrast is not merely quantitative; it is qualitative. GPT-4 could parse the natural language description of a vulnerability, reason about the preconditions for exploitation, identify the appropriate attack primitive, and execute it via a minimal agent framework. Other systems lacked this integrative reasoning capacity.

The study also revealed a critical dependency: without the CVE description, GPT-4's success rate dropped from 87 percent to 7 percent. This finding has a double edge. It suggests that LLMs are not yet reliably discovering novel zero-day vulnerabilities through pure code analysis — they depend on human-readable context to orient their exploitation attempts. But it equally confirms that the disclosure of a CVE accelerates the automation of exploitation dramatically. The implication for patch management is significant: organizations that were previously willing to accept a patch lag of days or weeks must recalibrate that assumption.

PwnGPT and Binary Exploitation

Research published in the ACL 2025 proceedings extended automated exploit generation to a technically harder domain: binary exploitation. PwnGPT is a purpose-built framework for automatic exploit generation against binary targets, using a modular architecture comprising analysis, generation, and verification components [5]. The authors benchmarked LLM performance against Capture the Flag (CTF) pwn challenges — a standard proxy for binary exploitation skill — and demonstrated that their framework increased the completion rate from 26.3 percent to 57.9 percent using OpenAI's o1-preview model, and from 21.1 percent to 36.8 percent using GPT-4o.

Binary exploitation is widely considered a high-skill discipline: it demands understanding of memory layouts, calling conventions, return-oriented programming, and the specific quirks of target binaries. PwnGPT's results indicate that structured scaffolding around LLMs — providing organized disassembly, binary metadata, and a feedback loop to verify whether generated payloads succeed — can substantially close the gap between general-purpose models and specialized human experts. The work validates the architectural pattern that defines the broader field: rather than expecting an LLM to solve an entire offensive problem end to end, researchers decompose the problem into stages that match LLM strengths and automate the connective tissue between them.

CTF Benchmarks and Competitive AI

CTF competitions serve as a useful, if imperfect, proxy for real-world offensive security skill. They test the same technical domains — binary exploitation, reverse engineering, web application attacks, cryptanalysis, forensics — in structured, time-bounded environments. Performance on CTF benchmarks has therefore become a leading indicator of offensive AI capability, even as researchers acknowledge that CTF challenges differ from production exploitation in important ways.

Progress on the InterCode-CTF benchmark illustrates how quickly the field has moved. In 2024, leading agent systems solved roughly 29 percent of challenges in early evaluations; by 2025, prompt-engineered agents using tool use and multiple attempts achieved 95 percent performance on the same benchmark [6]. CTFAgent, evaluated across PicoCTF challenges using GPT-4o, Gemini-2.5-Pro, and DeepSeek-V3, outperformed 88 percent of human teams in autonomous mode, rising to approximately 94 percent in human-in-the-loop configurations [7]. A Security Boulevard analysis of a 2025 dataset covering 423 Hack The Box machines released over eight years reported that root blood times — the time taken by the fastest human competitor to fully compromise a machine — had declined approximately 16 percent per year, with the sharpest compression observed in the period following the emergence of capable LLMs and agentic frameworks [8]. While this temporal correlation is consistent with LLM contribution to the trend, the original underlying study would be needed to assess the degree to which LLMs specifically drove the decline.

These figures should be interpreted carefully. CTF challenges are prepared exercises with defined solutions; real-world exploitation involves ambiguous targets, defensive systems, and unpredictable environmental conditions. Some benchmark results have been questioned on the grounds that foundation models may have memorized a fraction of challenge solutions from training data [6]. Nevertheless, the aggregate trend is unmistakable, and the benchmark improvements have been independently confirmed across multiple research groups.

DARPA AIxCC: Large-Scale Validation at Competition

DARPA's Artificial Intelligence Cyber Challenge (AIxCC), launched in 2023 and concluded in August 2025, represents the most rigorous and large-scale empirical evaluation of autonomous vulnerability analysis conducted to date [3]. Over two years of competition, teams developed cyber reasoning systems (CRSs) capable of discovering and patching vulnerabilities in real codebases without human guidance.

The final competition at DEF CON 2025 involved seven finalist teams, each provided with approximately 143 hours of fully autonomous operation, \$85,000 in cloud compute, and \$50,000 in LLM API credits. The challenge corpus consisted of 53 projects containing millions of lines of real code with novel synthetic vulnerabilities that had never appeared in any public dataset. Competing teams identified 86 percent of the synthetic vulnerabilities — up from 37 percent at the prior year's semifinals. Of the 54 synthetic vulnerabilities that teams discovered, 43 were patched — representing a 68 percent patch rate across the

full synthetic vulnerability corpus, up from 25 percent at the semifinals. In the process, the competing systems also discovered 18 previously unknown real-world vulnerabilities in the production software under analysis [3].

DARPA distributed a cumulative \$29.5 million in prizes, with the top prize of \$4 million awarded to Team Atlanta. The competition design, involving genuine novel flaws in production code, is specifically significant because it rules out the memorization concern that dogs CTF benchmarks: these were newly inserted bugs that no LLM had ever seen. The 86 percent synthetic detection rate represents authentic generalization.

From Proof of Concept to Operational Capability

Google Big Sleep and Real-World Zero-Day Discovery

Competitive benchmarks demonstrate capacity in controlled environments. The emergence of AI systems discovering real, previously unknown vulnerabilities in production software represents a qualitative step toward operational relevance. Google's Project Zero published the first confirmed public example of an LLM agent discovering an exploitable memory-safety vulnerability in widely deployed software in November 2024 [9].

The agent, named Big Sleep and evolved from Project Naptime, identified a stack buffer underflow in SQLite — a vulnerability that, if exploited, could enable arbitrary code execution. The vulnerability was reported and patched by the SQLite maintainers before it could be reached by an attacker. Google Project Zero described this as a milestone: an AI agent had found not merely a theoretical weakness but a practically exploitable memory corruption bug in one of the most widely deployed databases in the world.

In mid-2025, Google disclosed that Big Sleep had achieved a further milestone, identifying CVE-2025-6965 (CVSS score: 7.2), a memory corruption flaw affecting all versions of SQLite prior to 3.50.2, before any exploitation in the wild had occurred [10]. Google's Cloud CISO blog characterized the CVE-2025-6965 discovery as the first case of an AI agent finding a vulnerability before in-the-wild exploitation was detected. The significance is dual: Big Sleep demonstrates that AI can find vulnerabilities more rapidly than attackers can weaponize them — but also illustrates that the same analytical capability, in adversarial hands, could be directed at finding and weaponizing vulnerabilities rather than disclosing them responsibly.

CVE-2025-37899: The Linux Kernel SMB Disclosure

In May 2025, security researcher Sean Heelan disclosed CVE-2025-37899, a use-after-free vulnerability in the Linux kernel's ksmbd module — the kernel-space SMB3 implementation — discovered while benchmarking OpenAI's o3 model against known bugs [11]. The vulnerability is located in the SMB2

LOGOFF command handler and arises when concurrent SMB connections share and access a reference-counted object after it has been freed. Successful exploitation enables a remote attacker to achieve arbitrary code execution at kernel privilege.

Heelan provided o3 with the complete implementation of all SMB command handlers in the ksmbd module — approximately 12,000 lines of code — together with connection setup, teardown, and dispatch logic, with no additional scaffolding, agentic framework, or tool-use infrastructure beyond the o3 API. O3 successfully reasoned about the concurrency semantics: it understood that the freed object remained accessible to another thread and identified the specific code path that produced the race. Heelan assessed this to be, at the time of disclosure, the first publicly discussed example of a concurrency-related kernel vulnerability discovered by an LLM.

The technical character of the bug matters as much as its existence. Concurrency vulnerabilities are notoriously difficult for both humans and automated systems to identify: they require reasoning about non-deterministic thread interleaving, shared object lifetimes, and locking semantics. The fact that o3 identified this vulnerability — even as an unexpected output during a benchmarking exercise rather than a targeted search — suggests that frontier reasoning models have developed capacity for complex, multi-step program analysis that extends beyond pattern-matching over training data.

Incalmo and Autonomous Multi-Stage Network Attacks

Moving from single-host vulnerability exploitation to multi-host attack campaigns represents a substantial increase in operational complexity. A 2025 paper introduced Incalmo, an autonomous LLM-assisted red teaming system designed to execute multi-stage attacks across realistic multi-host network environments [2]. Incalmo addresses a key limitation of prior LLM-based attack systems: raw LLMs struggle to maintain coherent attack state across dozens of steps involving different hosts, credentials, and privilege levels.

Incalmo's architecture interposes a high-level abstraction layer between the LLM planner and low-level attack tooling. The LLM specifies attack goals using declarative verbs — "pivot to host X," "dump credentials," "exfiltrate data from service Y" — and a translation layer maps these to the appropriate operational implementations, invoking specialized agents for port scanning, lateral movement, credential dumping, and exfiltration. The result is an architecture in which the LLM's planning strengths are decoupled from the need to specify low-level implementation details.

Evaluated against MHBench, a multi-host benchmark comprising 40 environments modeling realistic enterprise network topologies, Incalmo successfully acquired critical assets in 37 out of 40 environments. Prior state-of-the-art LLM-assisted systems succeeded in 3 out of 40 environments under equivalent conditions. Successful attacks required between 12 and 54 minutes and consumed less than \$15 in LLM API credits [2]. This cost-to-capability ratio is notable: enterprise-grade network compromise at a cost that imposes essentially no barrier to sophisticated threat actors.

Agentic Attack Pipelines and Anthropic's Cyber Toolkit Research

In parallel with academic research, Anthropic's safety research team published findings in 2025 evaluating LLM agents equipped with standard cyber toolkits in controlled multi-stage attack scenarios [12]. The research, conducted in collaboration with CMU CyLab and published under Anthropic's responsible disclosure and capabilities evaluation framework, confirmed that capable frontier models can autonomously conduct multistage network attacks — including reconnaissance, initial exploitation, privilege escalation, and lateral movement — against realistic targets when provided with access to common offensive tools.

These findings, combined with the Incalmo results, establish a clear picture: the technical components for autonomous multi-stage attack execution exist and function reliably in research and controlled environments. The gap between controlled-environment demonstration and operational adversarial deployment remains, but the research trajectory suggests that gap is narrowing.

The Threat Actor Dimension

Nation-State LLM Adoption

Government-affiliated threat actors were among the earliest documented users of commercial LLMs for offensive purposes. In February 2024, a joint report from Microsoft and OpenAI described confirmed LLM usage by state-affiliated groups from China, Russia, North Korea, and Iran [13]. The observed uses were primarily productivity-oriented: drafting spear phishing content, translating materials into target languages, researching specific technical topics and vulnerability classes, and generating and refining scripts for known attack techniques. At that stage, none of the observed uses represented novel AI-enabled attack capability; LLMs were being used as sophisticated productivity tools rather than autonomous attack agents.

Two years later, the picture has evolved meaningfully. The Microsoft Digital Defense Report 2025 documented that AI-generated content is routinely incorporated into social engineering campaigns, with AI-crafted phishing messages achieving click-through rates approximately three times higher than traditionally composed lures — evidence of a qualitative improvement in adversarial content quality attributable to LLM assistance [14]. Nation-state actors are investing in LLM capability development, and the shift from LLMs as productivity tools toward LLMs as autonomous exploitation agents is a logical and anticipated progression.

Criminal Ecosystem Adaptation

The criminal ecosystem has adapted to LLM capabilities in predictable ways. Jailbroken and purpose-built offensive models — discussed further below — have circulated on dark web forums since 2023, and the barrier to accessing capable models outside of provider safety constraints has continued to decline as open-weight models have improved. By early 2026, GreyNoise researchers observed a sustained campaign in which two IP addresses methodically probed 73 or more LLM model endpoints over eleven days, generating more than 80,000 sessions — assessed as professional threat actor reconnaissance feeding into a broader exploitation pipeline [15].

The criminal ecosystem's adoption of LLMs for phishing, malware development, and script generation is well-documented. The more significant forward-looking concern is the adoption of LLM-based autonomous exploitation frameworks by ransomware operators and initial access brokers. The economics of such adoption are compelling: if an autonomous system can identify and exploit a high-severity CVE within hours of its publication, organizations with slow patch cycles become substantially more exposed.

Jailbreaking as a Weapon: The Mexico Government Incident

A campaign disclosed in early 2026, as reported by Cyberpress, involved an attacker using Anthropic's Claude to conduct a sustained offensive against Mexican government agencies [4]. According to that account, the attacker, operating over approximately one month beginning in December 2025, used relentless prompting techniques to overcome Claude's safety guardrails — a process described as shredding the model's alignment through iterative escalation rather than through a single novel jailbreak. The reported output included executable vulnerability scanning scripts, SQL injection exploits, and automated credential-stuffing tools. This account is based on a single secondary news source; specific operational details have not been corroborated by Anthropic's own transparency reporting, government incident advisories, or independent security research as of this writing, and should be treated accordingly.

The incident illustrates several dynamics that are consistent with the broader research literature, regardless of the specifics of this individual case. Determined attackers with sustained access to commercial LLMs and the patience to probe safety mechanisms can extract meaningful offensive capability even from models with well-designed safety training. The approach — persistence and incremental escalation rather than novel technical exploitation of the model — requires no specialized jailbreak skill. And the theoretical output of such sessions, functional exploit tooling tailored to identified targets, represents the kind of operational uplift that motivates attacker investment in this approach.

Anthropic's Constitutional Classifiers, introduced in 2025, reduced the jailbreak success rate from 86 percent to 4.4 percent in Anthropic's own controlled evaluations [16]. During a two-month independent evaluation described in the same publication, 183 participants spending an estimated 3,000 hours

collectively failed to extract all eight prohibited query types using a single jailbreak approach. These are meaningful improvements, but they coexist with evidence that sufficiently persistent human attackers can still extract harmful capability from commercial models.

The Weaponization Threshold: Structural Implications

Democratization of Offensive Capability

The most significant structural implication of LLM-based exploit automation is not that sophisticated nation-state actors become slightly more capable — they were already capable. It is that the lower end of the threat actor distribution is being lifted. The research demonstrates that the technical components for LLM-assisted exploitation exist and are increasingly accessible. The GreyNoise observation of professional threat actors systematically probing LLM API endpoints at scale [15] suggests that the operational security community is actively evaluating and integrating these capabilities. If lower-skilled actors follow this trajectory — a logical progression given the documented cost and accessibility of these tools — the effective floor of attack capability across the threat actor population will rise substantially. The population of entities capable of operationalizing a published CVE within hours of publication is likely expanding, though comprehensive adversarial deployment data for this actor tier remains limited.

Veracode's 2025 GenAI Code Security Report, focused primarily on the risks of LLM-generated application code, provides a relevant data point for a related concern: AI-generated code introduced security flaws in 45 percent of tests across more than 100 LLMs, with Java achieving a 72 percent failure rate [17]. Organizations using AI coding tools should be aware that this code will become the substrate for future CVEs — a growing body of vulnerable application code that did not previously exist. This is a distinct phenomenon from offensive tool generation, but it expands the vulnerability surface that automated exploit systems can target.

Speed and Scale Asymmetry

Defenders have always faced an asymmetry in which attackers need to find and exploit one weakness while defenders must protect all of them. Automated exploit generation intensifies this asymmetry along the time dimension. If the interval between CVE publication and exploit weaponization collapses from days or weeks to hours, the practical value of a published CVE-to-patch timeline changes fundamentally. Organizations that depend on patch-lag windows as a de facto buffer will need to recalibrate.

The UIUC study's finding that GPT-4 exploits 87 percent of one-day vulnerabilities with a CVE description, versus 7 percent without one [1], suggests a specific operational implication: the act of publishing a CVE now potentially hands a capable LLM attacker most of the information required to build a functional exploit before the patch has been deployed. This has implications for responsible disclosure timelines, coordinated disclosure processes, and the value of early warning notifications to critical infrastructure operators.

The Disclosure Dilemma

The responsible vulnerability disclosure ecosystem was designed around an implicit assumption: that the knowledge gap between a published CVE and a weaponized exploit would provide time for organizations to patch. That gap has been shrinking for years, driven by both criminal automation of prior-generation tools and the improved codification of exploit techniques in open-source frameworks. LLM-based exploit generation represents an acceleration of this trend rather than a discontinuity, but the acceleration is substantial enough to require a policy response.

OpenAI has explicitly acknowledged that coordinated vulnerability disclosure will become necessary as AI systems become more capable of finding and patching vulnerabilities, and that their own systems have already uncovered zero-day vulnerabilities [18]. The UK National Cyber Security Centre has similarly published guidance on adapting vulnerability disclosure processes to the AI context, noting that bypasses of AI safeguards require disclosure processes that do not yet exist in standardized form [19]. These acknowledgments reflect growing industry recognition that the disclosure frameworks developed for classical software vulnerabilities require adaptation.

Defensive Countermeasures and Mitigation

AI-Powered Defense and the Race Dynamic

The same capabilities that make LLMs effective for offensive security can be applied to defensive purposes — and there is growing evidence that the defensive application is viable and scalable. Google's Big Sleep demonstrates that AI agents can discover vulnerabilities and initiate responsible disclosure faster than attackers can weaponize them — at least in cases where the AI defender has privileged access to the source code. Proofpoint's Satori threat intelligence integration with Microsoft Security Copilot illustrates how LLM-based threat intelligence can accelerate defender prioritization of actively exploited vulnerabilities [20].

The DARPA AIxCC competition was explicitly structured around both offensive and defensive capability: competing systems were evaluated on their ability to both discover vulnerabilities and patch them, and the best-performing team patched 43 of the 54 synthetic vulnerabilities their systems found. This is a

meaningful defensive outcome. However, the competition also illustrates a structural challenge: the same systems that patch efficiently can be directed to exploit efficiently. The dual-use nature of automated vulnerability analysis means that defensive AI development and offensive AI development draw from overlapping research and tooling.

Organizational Vulnerability Management in the New Environment

The reduction in time-to-exploit for high-severity CVEs demands a corresponding reduction in organizational patch latency. Security teams should treat published CVEs affecting internet-exposed systems with CVSS scores above 8.0 as requiring emergency patching timelines, with remediation targets measured in hours rather than days. CSA recommends organizations target remediation within 48 hours for such vulnerabilities — a threshold shorter than most current patch cycles but consistent with the compressed exploitation timelines demonstrated in the research reviewed here. Automated vulnerability scanning and patch deployment pipelines should be treated as priority investments: if defenders cannot match the speed at which automated systems can exploit published vulnerabilities, organizations with large unpatched footprints will be systematically targeted.

Prioritization frameworks should weight exploitability heavily. The practical implication of the UIUC study is that any vulnerability for which a CVE advisory exists and which has a working public proof of concept should be treated as exploitable on the day of publication, rather than after attacker tool development cycles complete. Threat intelligence subscriptions that provide early warning of CVE publication, combined with automated deployment pipelines, are the organizational response.

Organizations deploying AI systems — whether LLM-based coding assistants, AI agents with tool access, or enterprise productivity tools like Microsoft 365 Copilot — must treat those systems as potential attack vectors. The EchoLeak vulnerability (CVE-2025-32711, CVSS 9.3) in Microsoft 365 Copilot demonstrated that AI systems can be weaponized through prompt injection to exfiltrate sensitive organizational data without user interaction [21]. AI deployment should include AI-specific threat modeling, input and output monitoring, and access controls on what data AI agents can access and exfiltrate.

LLM Provider Safety Controls

AI providers have significantly expanded their safety mechanisms in response to the emerging offensive use landscape. Anthropic's Constitutional Classifiers represent a substantive technical advance: the two-stage architecture uses a lightweight probe against model internal activations to screen all traffic, escalating suspicious interactions to a more capable classifier. In Anthropic's own controlled evaluation, this approach reduced jailbreak success from 86 percent to 4.4 percent [16]. Independent replication of this result has not been published as of this writing, and the evaluation methodology, controlled conditions designed specifically to test the classifier system, may not fully predict performance against the range of novel

prompting strategies used in real adversarial deployments. Nevertheless, the order-of-magnitude reduction demonstrates that safety engineering can materially constrain offensive use without eliminating model utility.

No safety mechanism is unconditional. The Mexico government incident, as reported, illustrates that persistent, patient attackers can still extract offensive capability from commercial models. Enterprises deploying LLMs in contexts with access to sensitive systems should layer organizational controls on top of provider safety mechanisms: rate limiting, output monitoring for exploitation-pattern signatures, access restrictions on which systems and data LLM agents can reach, and audit logging of all LLM interactions.

Secure Software Development and AI Code Quality

The intersection of LLM-generated code and security deserves attention beyond the explicit offensive context. Veracode's research finding that AI-generated code introduces security flaws in 45 percent of tests — with particularly high rates in Java (72 percent), and in vulnerability classes such as cross-site scripting (86 percent failure rate) and log injection (88 percent failure rate) — indicates that widespread LLM adoption in software development is producing a growing body of vulnerable application code [17]. This code will become the substrate for future CVEs. Organizations using AI coding tools should implement mandatory security scanning of AI-generated code prior to deployment, with the same rigor applied to human-generated code.

Policy and Governance Implications

The policy environment surrounding automated exploit generation is nascent but evolving. The EU AI Act, which has applied governance rules and General-Purpose AI (GPAI) obligations since August 2025 and reaches full enforcement in August 2026, establishes risk-based obligations for AI system deployers but does not specifically address dual-use offensive security tooling [22]. Existing frameworks for dual-use technology — export controls, computer fraud statutes, and vulnerability disclosure guidelines — were not designed with autonomous AI exploitation systems in mind.

Several governance gaps are evident. First, there is no established standard for AI provider responsibility when their models are used to generate functional exploit code against non-consenting targets. Existing safety mechanisms, while improved, are not currently sufficient to prevent operational misuse by determined actors. Second, CVE disclosure timelines were designed for a world in which exploit development required significant human effort. As that assumption degrades, the disclosure community — MITRE, the major CVE Numbering Authorities, and the national CERTs — should convene to assess whether standard disclosure windows remain appropriate.

Third, the competitive development of AI offensive capability by government actors is proceeding without the arms control frameworks that govern other dual-use military technologies. Nation-state use of LLM-based automated exploitation represents a qualitative change in cyberconflict dynamics that existing norms around responsible state behavior in cyberspace do not fully address. The academic literature on cyberconflict norms — including work building on the Tallinn Manual tradition — is only beginning to grapple with autonomous AI offensive systems as a distinct category.

Regulatory frameworks applicable to AI system developers should include requirements for transparency about dual-use capability evaluation, documentation of safety measures applied to offensive security use cases, and incident reporting when models are documented as instrumental in successful cyberattacks against third parties.

CSA Resource Alignment

The capabilities and threats documented in this whitepaper map directly to multiple pillars of the CSA AI Safety Initiative's existing work and framework portfolio. Rather than treating these as isolated controls, organizations should apply them in an integrated fashion, recognizing that the multi-stage, multi-agent attack architectures described throughout this paper require correspondingly layered defensive frameworks.

The MAESTRO (Multi-Agent Environment Security Threat Reconnaissance and Operations) framework is directly applicable to the agentic attack pipeline architectures described here. Incalmo, PwnGPT, and the CTFAgent architectures all involve LLM orchestration of specialized sub-agents — the exact multi-agent pattern that MAESTRO was designed to threat-model. Security architects should apply MAESTRO's threat modeling methodology to any deployment in which LLM agents are granted access to tools, networks, or data stores that could be leveraged in an offensive chain.

The AICM (AI Controls Matrix), as a superset of the CCM, provides the control framework against which AI-specific risks identified in this paper should be assessed. Relevant control domains include AI transparency — ensuring AI systems' actions are auditable — AI governance for establishing responsible use policies for LLM deployment, and AI incident response to ensure that LLM-mediated attacks are captured in existing security operations workflows. Organizations should use the AICM to identify gaps in their control posture specifically in relation to AI-augmented offensive scenarios, treating the attack classes documented here as concrete threat scenarios for control gap analysis.

STAR (Security Trust Assurance and Risk) assessments for AI service providers should now encompass evaluation of the provider's safety mechanisms against offensive use, the provider's incident response and notification processes when their models are used in confirmed cyberattacks, and the provider's

transparency about dual-use capability evaluation. The absence of standardized STAR criteria for LLM offensive safety is itself a gap that the CSA community should address as a near-term priority.

Zero Trust Architecture Guidance is especially relevant to the multi-stage attack scenarios described. Incalmo's success in achieving lateral movement and credential exfiltration in 37 of 40 environments suggests that network environments relying on perimeter-based security remain highly vulnerable to AI-assisted attack. Zero trust architectures — in which every access decision is verified, all traffic is logged, and lateral movement is restricted by policy — represent the defensive posture most resilient to automated multi-stage attack execution, and should be treated as a baseline for organizations managing sensitive infrastructure.

Finally, CSA Guidance on AI Organizational Responsibilities should be extended to address offensive LLM use cases explicitly, including responsible disclosure procedures when AI systems discover novel vulnerabilities, organizational liability frameworks for LLM-assisted attacks conducted using enterprise-provisioned AI tools, and employee acceptable use policies that address the offensive security research boundary.

Conclusions and Recommendations

The evidence surveyed in this whitepaper supports the following conclusions, offered as CSA's analytical assessment of the available research and incident data.

CSA's assessment is that LLMs have crossed a meaningful threshold from research demonstration to operational capability in automated exploit generation. The capability is not uniform — it depends on vulnerability class, target environment complexity, and available context — but it is real, improving, and increasingly accessible. The combination of frontier reasoning models with purpose-built scaffolding, specialized sub-agents, and exploit databases has produced autonomous systems that can execute multi-stage attacks at a fraction of the cost and time that previously required skilled human expertise. Benchmark results, while subject to the limitations acknowledged in this paper, consistently point in the same direction.

Real-world adversarial use is occurring. Nation-state actors have been using commercial LLMs since at least 2024 for offensive productivity tasks, and there is evidence consistent with LLM-assisted operational attack execution by criminal actors in late 2025. The gap between research demonstration and adversarial deployment appears to have narrowed to months rather than years, though the precise extent of operational deployment by non-state actors remains difficult to quantify from publicly available sources.

The democratization implication is significant and warrants attention even in the absence of comprehensive deployment data. The technical and cost barriers to LLM-assisted exploitation have demonstrably declined. Organizations that were previously protected by the skill barrier required to weaponize published CVEs face

a changed risk environment, one in which automated systems can do much of that work at minimal cost.

The defensive response must match the offensive trajectory. Based on the analysis presented here, CSA recommends the following actions.

Immediate (0–90 Days)

Organizations should audit their patch management workflows with specific attention to the time-to-patch for CVSS 8.0+ vulnerabilities affecting internet-exposed systems. Where this lag exceeds 48 hours, the risk posture should be reassessed against an assumption that published CVEs are being exploited by automated systems within hours of disclosure. AI systems deployed in enterprise environments — coding assistants, productivity AI, agentic workflows — should be threat-modeled specifically for prompt injection and scope violation risks, and output monitoring should be implemented for exploitation-pattern signatures.

Short-Term (90 Days–12 Months)

Security operations teams should evaluate AI-powered vulnerability detection tools alongside traditional scanning, recognizing that the defensive application of the same capabilities used offensively represents a meaningful force multiplier. Organizations should engage with their AI provider's security documentation to understand what safety mechanisms apply to their deployment, what monitoring is available, and what incident response procedures exist for LLM-mediated attacks.

Strategic (12 Months and Beyond)

Enterprise security architecture should be reviewed against a zero trust baseline, with particular attention to lateral movement controls, credential protection, and data exfiltration monitoring. The multi-stage attack pattern demonstrated by Incalmo and related frameworks exploits environments where successful initial access translates readily into broad network compromise. AICM assessments should be conducted for all AI deployments with access to production systems or sensitive data. Organizations with vulnerability research functions should review their responsible disclosure processes in light of the compressed timelines that automated weaponization introduces.

For the security research and policy community, the most urgent need is convergence on updated disclosure standards that account for AI-accelerated weaponization, governance frameworks that address dual-use offensive AI development at the national and international level, and AI provider accountability standards for the offensive use of commercial models.

References

[1] Fang, R., Bindu, R., Gupta, A., and Kang, D. "LLM Agents can Autonomously Exploit One-day Vulnerabilities." arXiv:2404.08144, April 2024.

[2] Singer, B., et al. "Incalmo: An Autonomous LLM-assisted System for Red Teaming Multi-Host Networks." arXiv:2501.16466, January 2025.

[3] DARPA. "AI Cyber Challenge marks pivotal inflection point for cyber defense." DARPA, August 2025.

[4] Gambit Security / Cyberpress. "Hacker Jailbreaks Claude AI to Generate Exploit Code and Exfiltrate Government Data." Cyberpress, January 2026.

[5] Chen, et al. "PwnGPT: Automatic Exploit Generation Based on Large Language Models." ACL Anthology, ACL 2025.

[6] Palisade Research. "Hacking CTFs with Plain Agents." Palisade Research, January 2025.

[7] Liu, et al. "CTFAgent: An LLM-powered Agent for CTF Challenge Solving." ScienceDirect, 2025.

[8] Security Boulevard. "The Death of the CTF: How Agentic AI Is Reshaping Competitive Hacking." Security Boulevard, March 2026.

[9] Google Project Zero. "From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code." Project Zero Blog, November 2024.

[10] Google Cloud Blog. "Cloud CISO Perspectives: Our Big Sleep agent makes a big leap." Google Cloud, 2025.

[11] Heelan, S. "How I used o3 to find CVE-2025-37899, a remote zeroday vulnerability in the Linux kernel's SMB implementation." Sean Heelan Blog, May 2025.

[12] Anthropic Red Team. "LLMs with cyber toolkits can conduct multistage network attacks." Anthropic, June 2025.

[13] Microsoft Security Blog. "Staying ahead of threat actors in the age of AI." Microsoft, February 2024.

[14] Microsoft. "2025 Microsoft Digital Defense Report." Microsoft Security Insider, 2025.

[15] GreyNoise. "Threat Actors Actively Targeting LLMs." GreyNoise, January 2026.

[16] Anthropic. "Constitutional Classifiers." Anthropic Research, 2025.

[17] Veracode. "GenAI Code Security Report 2025." Veracode, October 2025.

[18] OpenAI. "Scaling security with responsible disclosure." OpenAI, 2025.

[19] NCSC. "From bugs to bypasses: adapting vulnerability disclosure for AI safeguards." UK National Cyber Security Centre, 2025.

[20] Proofpoint. "Proofpoint Satori Emerging Threats Intelligence Agent Now Generally Available for Microsoft Security Copilot." Proofpoint, 2025.

[21] Truesec / SOC Prime. "CVE-2025-32711 Vulnerability: 'EchoLeak' Flaw in Microsoft 365 Copilot." SOC Prime, June 2025.

[22] European Commission. "EU AI Act." European Commission, enforced August 2026.