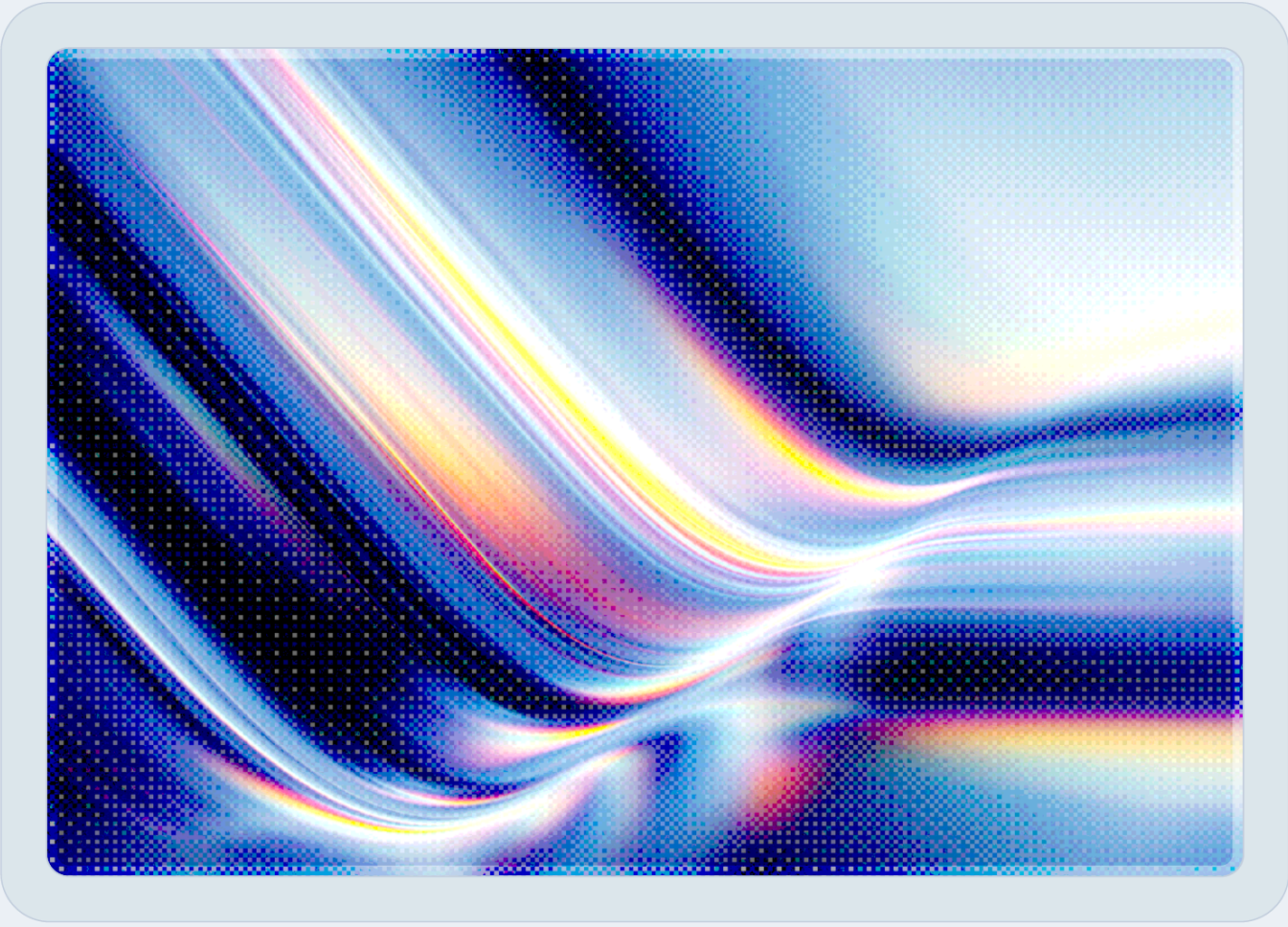
CSAI Foundation | Cloud Security Alliance

Sub-Frontier AI Models Can Now Find Zero-Days

Capability Democratization Lowers the Barrier for Offensive AI Vulnerability Discovery

2026-05-12

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

- When University of Illinois Urbana-Champaign researchers published the first systematic study of LLM agent performance against real-world CVEs in April 2024, only GPT-4 could autonomously exploit real-world vulnerabilities, achieving an 87% success rate on disclosed but unpatched CVEs while every other tested model — including GPT-3.5, open-source LLMs, and conventional scanners — scored zero [1]. That frontier-only bottleneck no longer holds.
- Subsequent research and documented operational deployments indicate that vulnerability discovery and exploitation capability has spread meaningfully down the model hierarchy: smaller, cheaper, and in some cases open-weight models now show material offensive capability that was negligible as recently as eighteen months ago. Systematic multi-model benchmark studies of this transition are limited, but the trajectory is evident in the operational record.
- Google's Project Zero Big Sleep agent found a previously unknown, exploitable stack buffer underflow in SQLite in November 2024 — described by the team as the first public example of an AI agent discovering a genuine, in-production zero-day in widely deployed software [2][7].
- The economics of AI-assisted exploitation have shifted dramatically: the UIUC study estimated an average exploitation cost of $8.80 per vulnerability using GPT-4 versus an estimated $25 per vulnerability for a skilled human researcher, and subsequent model cost reductions have pushed this figure further downward [1].
- The speed advantage is compounding independently of model capability: industry data indicates that 28.3% of CVEs were exploited within 24 hours of public disclosure in Q1 2025, a rate that narrows the defender's patch window to the point where conventional vulnerability management timelines are functionally obsolete [3].
- Google's Threat Intelligence Group identified structural indicators in a recovered zero-day exploit consistent with AI-assisted authorship — the first reported case suggesting that capability democratization has moved from research demonstration to adversarial deployment [4][5].

Background

For most of computing history, the discovery of exploitable software vulnerabilities required a narrow combination of skills: deep knowledge of the target software's architecture, familiarity with memory layouts and compiler behavior, fluency in assembly or low-level language concepts, and the patience to manually trace execution paths through hundreds of thousands of lines of code. This expertise barrier functioned, imperfectly but meaningfully, as a natural access control on who could find and weaponize zero-day vulnerabilities. Nation-state intelligence agencies, well-resourced criminal organizations, and elite security research teams sat on one side of that barrier; most threat actors did not.

AI language models, as they became capable of reasoning about code, began to erode that barrier. The first rigorous evidence arrived in April 2024, when Daniel Kang and colleagues at UIUC published a controlled study testing LLM agents against fifteen real-world, reproducible vulnerabilities — a dataset spanning website vulnerabilities, container escape issues, and library-level flaws, most rated high or critical severity [1]. Their methodology was specific: they configured each model with tools that simulated a competent security researcher's environment and gave it access to the relevant CVE description alongside the target. GPT-4 successfully exploited 13 of the 15 vulnerabilities, an 87% success rate. Every other model in the evaluation — GPT-3.5, open-source alternatives, and commercial vulnerability scanners including Metasploit and OWASP ZAP — achieved a success rate of zero. At that moment, the capability was real but concentrated: it existed effectively at a single point on the capability frontier.

A second finding from the same study proved equally consequential for projecting capability trajectories. When the researchers removed the CVE description from GPT-4's context, its success rate fell from 87% to 7% — a collapse that suggested the model was not yet performing genuine generalization from code alone but was applying pattern-matching informed by structured vulnerability metadata. That finding carried an implicit warning: it bounded GPT-4's current capability while also pointing toward the trajectory that more capable models would follow as they learned to reason about code without scaffolding. The question was how long the frontier-only concentration would persist.

Security Analysis

The Frontier Bottleneck Is Breaking Down

The observation that only the most capable models could find and exploit vulnerabilities rested on a specific assumption: that the task demanded reasoning abilities that smaller models lacked entirely. The operational record through 2025 challenges that assumption. While systematic multi-model benchmark studies of this transition are limited, evidence from documented deployments and the trajectory of model cost reductions suggests that smaller and open-weight models have demonstrated non-trivial capability on subsets of the vulnerability space — particularly for vulnerabilities with simpler exploitation patterns or those where prior art is well-represented in training data.

This is not a marginal distinction. Even a model that succeeded on, say, 15–20% of vulnerabilities rather than GPT-4's 87% is still a tool that scales in ways human attackers cannot. A threat actor operating ten parallel agentic workflows with a sub-frontier model may achieve aggregate throughput on opportunistic scanning and exploitation that exceeds what a skilled researcher could accomplish manually. The relevant unit of analysis for defenders is not individual model capability but the system-level productivity of an attacker who can run many agents simultaneously at low cost against many targets.

The economic dimension reinforces the trend. Open-weight models of the size now showing meaningful vulnerability research capability can be run on mid-range gaming GPUs (24GB VRAM class hardware) or accessed via commodity inference APIs at prices far below those of frontier API access. The $8.80 per-vulnerability figure from the 2024 UIUC study reflects frontier API pricing at the time of the study; open-weight models running locally eliminate per-query API costs for operators with the infrastructure to deploy them [1]. That cost structure is accessible to a broader range of threat actors than frontier API subscriptions.

Zero-Day Discovery Has Crossed the Operational Threshold

The distinction between finding known vulnerabilities and discovering previously unknown ones has sharpened considerably. The UIUC 2024 study demonstrated exploitation of disclosed vulnerabilities — a valuable capability, but one bounded by the public CVE database. The more significant threshold is discovering vulnerabilities that no one has previously identified.

Google Project Zero's Big Sleep agent crossed that threshold in November 2024. The agent, which augments AI reasoning with specialized security research tooling and sandboxed execution environments, discovered a stack buffer underflow in SQLite that was reproducible, exploitable, and entirely unknown prior to the agent's investigation [2][7]. Project Zero described this as the first public

instance of an AI agent finding a genuine zero-day in production software with real-world deployment scale. SQLite is embedded in an estimated one trillion active installations globally [6], a fact that underscores what zero-day discovery capability means when applied to the most widely deployed software components.

The follow-on implication is that the AI vulnerability discovery capability relevant to defenders is not only the ability to automate exploitation of known CVEs — it is the potential to identify and weaponize vulnerabilities that exist in the wild but have never been catalogued. The traditional assumption that defenders who patch disclosed CVEs are adequately protected does not account for AI systems discovering and exploiting vulnerabilities that exist in deployed software before any patch has been written or any disclosure has occurred.

Exploitation Timelines Have Collapsed Independently

The democratization of AI vulnerability discovery compounds a separate trend that was already stressing conventional vulnerability management: the collapse of the exploitation timeline. Industry data from Q1 2025 shows that 28.3% of CVEs were exploited within 24 hours of public disclosure, a figure that describes a threat environment where the traditional model — disclose, notify vendors, allow patch development, push patches, wait for enterprise deployment — fails at the first link in the chain [3]. Exploits now routinely arrive before patches, and the window between CVE publication and active exploitation has compressed to hours in many cases.

AI-assisted exploitation accelerates this dynamic. Where a human attacker working from a CVE description might require days or weeks to develop a working exploit, an AI agent operating on the same description can in principle produce an exploit candidate within minutes. The UIUC study documented this concretely: given a CVE description and a target environment, GPT-4-era models could work through the exploitation logic within the timeframe of a single agent session. At the sub-frontier tier, where models may require more iterations or achieve lower success rates on any given vulnerability, parallel deployment across many instances can offset reduced per-instance capability with increased throughput.

The net effect is that the two trends — broader access to AI vulnerability discovery capability, and compressed exploitation timelines — reinforce each other in ways that are additive for attackers and compounding for defenders. A defender who has deployed adequate patching infrastructure for a 72-hour response window may find that window is no longer sufficient when the relevant exploits are being developed and distributed at machine speed.

Threat Actor Adoption Is Documented, Not Hypothetical

The transition from research demonstration to documented adversarial use has now occurred. Google's Threat Intelligence Group identified, in their 2025 reporting, a zero-day exploit bearing structural characteristics consistent with AI authorship — code organization, commenting patterns, and error handling logic that diverged from established human exploit author conventions [4][5]. The structural indicators pointed to AI-assisted development rather than AI acting as the primary author; the actor planned mass exploitation of the vulnerability before GTIG's proactive detection and coordinated disclosure disrupted the campaign.

Separately, nation-state threat clusters linked to China and North Korea have been documented using AI assistance across the attack chain, including for vulnerability enumeration and proof-of-concept development, at operational tempos that would be impractical without automation [4][8][9]. These are not experimental programs; they represent production-grade integration of AI tooling into adversarial operations at scale. The pattern observed in these nation-state operations — AI used not as a primary attack tool but as an accelerator and productivity multiplier across the full kill chain — is precisely the application model where sub-frontier models become viable: high-volume, lower-precision tasks that aggregate into operational advantage.

What "Democratization" Actually Means for Defenders

The term democratization can obscure the specific mechanism of risk. The concern is not primarily that unskilled attackers will gain the ability to discover zero-days comparable to an elite nation-state research team — the capability gap at the frontier remains significant. The concern is that the floor of capability accessible to moderately resourced threat actors has risen substantially, and continues to rise with each model generation.

A threat actor who could previously only deploy commodity malware using known techniques and publicly available exploit frameworks can now plausibly operate AI-assisted scanning and exploitation workflows at scale. A criminal group that would historically have lacked the expertise to capitalize on a newly disclosed critical vulnerability can now deploy LLM agents to attempt exploitation within hours of publication. The change is not categorical — sophisticated attackers becoming unsophisticated, or vice versa — it is distributional: more actors operating at higher capability levels than the historical distribution would predict.

For enterprise security teams, this means the population of actors capable of finding and exploiting vulnerabilities in their environments is larger than it was two years ago, and is growing with each model release cycle. Threat modeling assumptions calibrated to state-level-only zero-day exploitation no

longer hold for organizations facing criminal ransomware groups, hacktivist collectives, or nation-state-adjacent contractors, all of which now have plausible access to AI-assisted exploitation tooling.

Recommendations

Immediate Actions

- **Reassess vulnerability prioritization SLAs against current exploitation timelines.** If your organization's patch SLA for critical vulnerabilities is measured in days or weeks, it is calibrated for a threat environment that is no longer complete — exploitation capability has expanded to a broader population of actors than that model assumed. The documented pattern of same-day exploitation for disclosed CVEs should drive SLA review for CVSS 9.0+ vulnerabilities to hours, not days, with automated patch deployment workflows where feasible.
- **Audit external attack surface exposure for unpatched high-severity vulnerabilities.** Given the documented AI-assisted scanning and exploitation capability, publicly accessible services with unpatched critical vulnerabilities face an elevated automated exploitation risk. Emergency patching cadences for externally reachable systems should be treated as a standing operational priority rather than a response to specific threat intelligence.
- **Review AI security tooling subscriptions and access controls.** Organizations that have deployed or are evaluating AI-assisted penetration testing tools should ensure that access to those tools is governed by the same controls applied to other privileged security tooling, and that their use is logged and auditable.

Short-Term Mitigations

- **Implement compensating controls for software components with high-population deployment footprints.** Libraries and frameworks embedded in large numbers of applications — SQLite, OpenSSL, and similar foundational components — represent attractive targets for AI-assisted zero-day research because the exploitation surface is large and the payoff is disproportionate. Where runtime protection layers (memory-safe wrappers, sandboxing, privilege isolation) can be applied without operational impact, they reduce the exploitability of discovered vulnerabilities independent of patching.
- **Enhance monitoring for early indicators of AI-generated exploitation activity.** GTIG has documented distinguishing characteristics of AI-authored exploits, including regularized code structure, educational commenting, and hallucinated metadata (e.g., fictitious CVE

numbers or nonexistent function signatures inserted by the AI). Security operations teams should incorporate these behavioral signatures into detection rulesets and threat hunting playbooks, and subscribe to threat intelligence feeds that specifically cover AI-assisted offensive operations.

- **Evaluate AI-assisted defensive tooling to close the capability gap.** The same research demonstrating AI-assisted offensive capability also establishes that AI models can operate as defenders — identifying vulnerabilities in first-party code before attackers can find them, validating patch effectiveness, and accelerating triage of disclosed CVEs. Organizations without an AI-augmented security engineering practice are increasingly operating asymmetrically against adversaries who have one.

Strategic Considerations

The architecture of software supply chains amplifies the democratization risk described in this note. Foundational open-source libraries are embedded in enormous populations of dependent software, often without the owning organizations' knowledge of which versions they run. AI-assisted zero-day discovery targeted at these foundational components can produce vulnerabilities that affect millions of applications simultaneously, creating exploitation opportunities that scale beyond what targeted research against individual products would yield. Software bill of materials (SBOM) programs and continuous dependency monitoring are not merely governance hygiene; they are a prerequisite for managing the exploitability of the supply chain under conditions of AI-accelerated vulnerability discovery.

Longer term, the trajectory of sub-frontier model capability suggests that the threat landscape will continue to expand as each successive model generation pushes more capability below the cost and access thresholds relevant to a broader population of threat actors. Security strategies that depend on capability concentration at the frontier — the assumption that only well-resourced adversaries can find zero-days — are structurally misaligned with the direction of this trend. Defense programs built on continuous exposure reduction, compensating runtime controls, and AI-assisted vulnerability detection in first-party code are better positioned to remain effective as the capability distribution continues to shift.

CSA Resource Alignment

This research note connects to several areas of the Cloud Security Alliance's AI and cloud security frameworks.

The AI Controls Matrix (AICM) v1.0 [10] establishes the control architecture relevant to organizations deploying AI systems and consuming AI-generated security outputs. Its coverage of AI supply chain security and model provider accountability is directly relevant to organizations evaluating AI-assisted penetration testing and vulnerability management tooling: the same controls that govern what AI systems can do within an organization's environment apply to the offensive AI tools that the threat actors documented in this note are deploying against it.

CSA's MAESTRO threat modeling framework for agentic AI [11] provides the analytical structure most directly applicable to the autonomous exploitation agent pattern described here. MAESTRO's treatment of agentic autonomy, tool use, and multi-step reasoning maps directly onto the attack chain — target identification, CVE analysis, exploit development, and exploitation — that AI-assisted threat actors are operationalizing. Security teams using MAESTRO to model their own agentic AI deployments should apply the same framework to adversarial agentic workflows as a threat modeling exercise.

The STAR for AI program's risk assessment methodology supports the due diligence process for AI-assisted security tooling procurement. As organizations evaluate AI vulnerability scanning and penetration testing products, the STAR framework provides a structured basis for assessing whether vendor security claims are substantiated and whether the tooling's operational posture is consistent with the access it requires.

CSA's Zero Trust guidance is relevant to the compensating control layer this note recommends: runtime privilege isolation, network segmentation, and least-privilege access for the foundational software components most likely to be targeted by AI-assisted zero-day research reduce the blast radius of successful exploitation independent of patching velocity.

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

[3] VulnCheck. "Exploitation in 2025: Q1 Trends and Observations." VulnCheck, 2025.

[4] Google Threat Intelligence Group. "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools." Google Cloud Blog, 2025.

[5] SecurityWeek. "Google Detects First AI-Generated Zero-Day Exploit." SecurityWeek, 2025.

[6] SQLite. "Most Widely Deployed and Used Database Engine." SQLite.org.

[7] The Hacker News. "Google's AI Tool Big Sleep Finds Zero-Day Vulnerability in SQLite Database Engine." The Hacker News, November 2024.

[8] Mandiant / Google Cloud. "M-Trends 2025." Mandiant, 2025.

[9] Palo Alto Networks Unit 42. "Can AI Attack the Cloud? Artificial Intelligence Exploiting Initial Access Vulnerabilities." Palo Alto Networks, April 2026.

[10] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.0." Cloud Security Alliance, 2024.

[11] Cloud Security Alliance. "MAESTRO: Agentic AI Threat Modeling." Cloud Security Alliance, 2025.