ultraviolet

THREAT ADVISORY SPECIAL REPORT

Dark AI

Services Performed By:
UltraViolet Cyber TIDE Team
tide@uvcyber.com

Published Date:
April 22, 2026
TLP:GREEN

Executive Snapshot

The weaponization of artificial intelligence by threat actors has moved from theoretical risk to operational reality, with purpose-built malicious LLMs, AI-driven phishing and deepfake campaigns, and fully autonomous attack chains now targeting enterprise infrastructure at machine speed. Compounding external threats, shadow AI has become one of the fastest-growing insider risks, as employees deploy unsanctioned tools like Ollama on corporate hardware — often exposing unauthenticated inference endpoints to the public internet without security team awareness. Researchers have identified over 175,000 exposed Ollama instances worldwide, and the first documented LLMjacking marketplace has already demonstrated how attackers systematically discover, validate, and monetize unauthorized access to enterprise AI compute. With over 75% of organizations now reporting shadow AI as a definite or probable problem, internal ownership of AI security controls still contested in nearly three-quarters of enterprises, and AI-specific budget allocations remaining dangerously thin, the gap between AI adoption velocity and security readiness represents one of the most consequential unmanaged risks on the modern enterprise attack surface.

Recommended Action Items:

- **Expand asset discovery to include AI infrastructure.** Traditional vulnerability scanners do not prioritize AI-specific services. Security teams should extend scanning to cover default ports used by tools like Ollama (11434), vLLM (8000), LM Studio (1234), and Gradio (7860), and deploy fingerprinting capabilities that can identify inference endpoints, model types, and API signatures across cloud, on-premises, and developer workstation environments.

- **Enforce authentication and network isolation on all inference endpoints.** Ollama and similar frameworks ship without native authentication, and exposing them externally requires only a single configuration change. Organizations should mandate reverse proxy layers with strong authentication (OAuth, API keys, or SSO), restrict AI services to internal networks via firewall policy, and treat any internet-facing inference endpoint as a critical finding during routine security assessments.

- **Establish a formal AI governance framework with clear ownership.** The current ambiguity over whether AI security falls under the CISO, CTO, or data science leadership creates exploitable gaps. Organizations should designate a single accountable owner for AI security controls, publish an acceptable-use policy covering both cloud-hosted and locally deployed models, and implement DLP controls that monitor for sensitive data — proprietary code, PII, and strategic documents — flowing into unsanctioned AI services.

- **Provide sanctioned, secure AI alternatives that meet workforce productivity demands.** Shadow AI proliferates because employees face productivity pressure and corporate-approved tools are either unavailable, too slow, or too restrictive. Deploying enterprise-licensed AI platforms with zero-data-retention guarantees, SSO integration, and audit logging removes the incentive for workarounds, while offering managed local inference options for highly regulated environments ensures sensitive data never leaves the corporate boundary.

A TLP:AMBER version of this special report with additional information on this topic has been made available to UVCyber Customers and Partners.

TIDE Team Analysis

The convergence of artificial intelligence and cyber threats has produced a category of risk that demands immediate executive attention. Over the past three years, threat actors have moved from experimenting with generative AI to embedding it as a core execution engine within their attack toolchains. The average cost of a data breach reached $4.4 million in 2025, with phishing remaining the primary intrusion vector at roughly 60% of incidents, now delivered with unprecedented realism using AI-generated content. What the industry broadly refers to as "Dark AI" — malicious large language models, unsanctioned local deployments, and weaponized open-source inference tools — has matured from a theoretical concern into a measurable, operational threat to enterprise infrastructure. Security leaders who fail to account for this shift risk defending against yesterday's adversary while today's is already inside the perimeter.

In June 2023, a tool called WormGPT appeared on underground forums, built on the open-source GPT-J model and marketed as a ChatGPT alternative for criminal use. Unlike mainstream generative models with guardrails to block harmful outputs, tools like WormGPT and FraudGPT were explicitly designed for cybercrime — crafting spearphishing emails, generating undetectable malware, and facilitating digital impersonation. FraudGPT was offered on subscription ranging from $200 per month to $1,700 per year, with over 3,000 confirmed sales reported by mid-2023. While many of these early tools proved short-lived or fraudulent themselves, they demonstrated a critical principle: because foundational LLMs are open source, anyone with sufficient knowledge can fine-tune them into bespoke offensive tools, and the barrier to doing so continues to fall.

The threat has since evolved well beyond crude dark-web chatbots. In 2025, there was a sharp increase in AI-generated phishing and deepfake-enabled social engineering, with threat actors routinely using LLMs to craft convincing, tailored content at massive scale. Industry surveys reported that 85% of organizations experienced at least one deepfake-related incident in the past year. Attackers are now combining generative models with agentic automation — systems that can autonomously plan, probe external attack surfaces, chain exploits, and adapt to defender responses in real time. Security experts predict that offensive autonomous AI will emerge as a mainstream threat, with attackers deploying fully automated phishing, lateral movement, and exploit development. The shift from human-in-the-loop attacks to machine-led operations represents a fundamental change in the speed and scale at which adversaries can operate.

Within the enterprise itself, a parallel and often overlooked risk has taken root: shadow AI. Separate from the malicious use of AI tools that is Dark AI, Shadow AI refers to the use of AI tools without approval, visibility, or monitoring. More than three in four organizations now cite shadow AI as a definite or probable problem, up from 61% in 2025 — a 15-point year-over-year increase and one of the largest shifts in recent survey data. Employees bypass sanctioned platforms in favor of faster, free alternatives — personal ChatGPT sessions, local model installations, and unmanaged agents — driven by productivity pressure and tight deadlines. Recent surveys show over 75% of knowledge workers using generative AI at work, with 46–60% engaging in risky unauthorized behaviors, and shadow AI incidents now adding approximately $670,000 in extra remediation costs per breach. The result is a sprawling, invisible attack surface that traditional security controls were never designed to detect.

Tools like Ollama sit at the center of this blind spot. Ollama is an open-source framework that lets users run LLMs locally on personal or corporate hardware with minimal technical effort, and it has surpassed 155,000 stars on GitHub. A joint investigation by SentinelOne and Censys revealed 175,000 unique Ollama hosts exposed across 130 countries, operating entirely outside the guardrails and monitoring systems that platform providers implement by default. The project does not natively support authentication, and researchers at Wiz have noted that exposing the API to the internet requires only a trivial configuration change. Multiple critical vulnerabilities have been discovered in Ollama, including denial-of-service bugs, authentication bypass flaws, and remote code execution vectors. A

developer experimenting on a workstation can inadvertently expose an unauthenticated inference endpoint to the entire internet — and most enterprise vulnerability scanners will never flag it.

The consequences of this exposure are no longer hypothetical. Between December 2025 and January 2026, Pillar Security documented Operation Bizarre Bazaar, the first systematically attributed LLMjacking campaign, in which threat actors scanned for exposed Ollama instances and other AI endpoints, validated access, and resold it through a commercial marketplace at 40–60% discounts. In these LLMjacking attacks, threat actors use victims' electricity, bandwidth, and compute to generate spam, produce malware content, and resell access to other criminals. Nearly half of exposed hosts were configured with tool-calling capabilities that enable them to execute code, access APIs, and interact with external systems — meaning a compromised endpoint can serve as a pivot point into broader enterprise infrastructure. What security teams dismiss as a "dev tool" is, to an attacker, a privileged interpreter with network access.

The organizational challenge is structural, not merely technical. Seventy-three percent of organizations report internal conflict over ownership of AI security controls, and while 91% added AI security budgets in 2025, more than 40% allocated less than 10% of that budget to AI-specific security measures. The people who understand the data — developers and data scientists — and the people who secure the infrastructure — the CISO's team — often operate in separate worlds, creating a critical blind spot. Ollama defaults to port 11434; vLLM typically runs on 8000; LM Studio uses 1234. These are not ports that traditional scanners prioritize, and without deliberate effort to fingerprint AI-specific services, security teams cannot protect what they cannot see.

Security leadership must treat dark AI and shadow AI deployments as first-order infrastructure risks, not edge cases. This requires updating threat models to assume local LLM deployments exist anywhere developers work, extending asset discovery to include AI-specific ports and API signatures, enforcing authentication and network isolation on all inference endpoints, and establishing clear governance that balances productivity with security rather than driving usage underground. The urgent question is no longer whether the organization is running AI — it is where AI is running that the organization does not know about. The window between an unmanaged Ollama instance going live and an attacker discovering it is measured in hours, not weeks. Organizations that close that gap proactively will be positioned to harness AI's advantages; those that do not will find their own infrastructure weaponized against them.

Why It Matters

The trajectory of AI-enabled threats has followed a clear and accelerating arc. In mid-2023, the first malicious LLMs appeared as crude, short-lived experiments — tools like WormGPT and FraudGPT that were sold on underground forums for a few hundred dollars a month and often disappeared within weeks. They were rudimentary, frequently overhyped by their creators, and in some cases outright scams. But they proved a concept that fundamentally altered the threat landscape: open-source model weights could be fine-tuned for offensive purposes by anyone with modest technical skill, and the safety guardrails built into commercial AI platforms were trivially easy to circumvent by building outside them entirely. By 2025, the ecosystem had matured dramatically. Threat actors moved beyond standalone chatbots to integrate generative AI directly into their operational toolchains — automating reconnaissance, generating polymorphic malware variants that evade signature-based detection, and producing deepfake audio and video convincing enough to impersonate executives in real-time calls. The barrier to launching sophisticated, targeted campaigns dropped to near zero, and attack volume scaled accordingly.

The present moment marks a transition from AI-assisted attacks to AI-autonomous operations. Agentic AI systems — models capable of planning, executing, and adapting multi-step actions without human intervention — are now being embedded into adversary frameworks that can probe attack surfaces around the clock, chain exploits across environments, and pivot laterally through infrastructure at speeds no human operator could match. Simultaneously, the internal threat surface has expanded in ways most security programs have not accounted for. The explosion of locally hosted inference tools like Ollama has created a distributed, ungoverned layer of AI compute sitting inside and alongside enterprise networks, often deployed by well-intentioned employees who have no idea they have exposed a privileged, unauthenticated API to the internet. Operation Bizarre Bazaar demonstrated in early 2026 that criminal ecosystems have already industrialized the discovery and monetization of these endpoints — scanning, validating, and reselling access through commercial marketplaces with the same operational efficiency as legitimate SaaS businesses.

Looking ahead, these trends will converge and compound. As open-weight models continue to improve in capability and shrink in size, deploying powerful AI locally will become easier, cheaper, and harder to detect. Fine-tuned offensive models will grow more specialized — purpose-built for specific industries, attack types, or defensive evasion techniques — and the criminal marketplace for AI-as-a-service will mature alongside them. Data poisoning attacks against enterprise training pipelines, adversarial manipulation of agentic AI systems, and the exploitation of Model Context Protocol servers as lateral movement vectors are all emerging threat categories that most organizations have not yet incorporated into their risk models. The organizations that treat AI security as an afterthought or a niche technical concern will find themselves structurally disadvantaged — not in some distant future, but within the next twelve to eighteen months. The window to get ahead of this threat is narrowing, and the cost of inaction is compounding with every unsanctioned deployment, every unmonitored endpoint, and every month that passes without a coherent governance framework in place.

How to Respond

- Strictly adhere to cybersecurity Fundamentals and ensure all personnel undergo annual phishing and social engineering training. Speak with your UltraViolet Cyber TAM Representative to schedule a live phishing engagement.
- Ensure all users, especially users with elevated permissions, are aware of the risks that Dark AI tools can pose.
- Verify that the Acceptable Use Policy has language that covers Shadow AI and IT.
- Perform annual tech refresh reviews to gain a holistic understanding of your infrastructure. Speak with your UltraViolet Cyber TAM Representative to schedule a RedTeam or PurpleTeam engagement to gain insight into the vulnerabilities in your environment.

What UltraViolet Cyber is Doing

- Testing new open source LLM models as they come out and verifying if they can be used for malicious purposes.
- Proactively enabling custom detections based on the collected artifacts, tactics, techniques, and procedures identified in this activity.
- Performing hypothesis driven threat hunts based on threat actor behavior and artifacts. UVCyber customers will be informed of the results through secure channels.
- Parsing available victim dump data for any social, financial, business, or technical relations to UVCyber Clients and partner organizations.
- Aggregating threat intelligence from myriad sources and applying the most up-to-date knowledge to proactive threat hunting and response.

About UltraViolet Cyber

UltraViolet Cyber is a leading tech-enabled managed security services provider, delivering unparalleled cybersecurity expertise that fills technology and talent gaps across Global 2000 and Federal Government customers. Founded and operated by security practitioners from the national intelligence community, UltraViolet Cyber connects offensive security, application security, detection and response, and security engineering to deliver a differentiated approach to cybersecurity operations. Transforming customers' security programs, UltraViolet Cyber's flagship security-as-a-service solution, UV Lens, removes complex operational silos, replacing them with integrated security capabilities. UltraViolet is headquartered in McLean, Virginia with technology centers across the world.

443.351.7630 / info@uvcyber.com | UltraViolet Cyber | @uv_cyber

---