ultraviolet

THREAT ADVISORY

JADEPUFFER: AI-Driven Ransomware Attacks

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

Published Date:
July 08, 2026
TLP:GREEN

Executive Snapshot

Researchers at Sysdig's Threat Research Team have documented what is assessed to be the first known case of a ransomware operation run entirely by an autonomous large language model, with no human operator directing any phase of the attack. The actor, designated JADEPUFFER, moved through reconnaissance, credential harvesting, lateral movement, data encryption, and extortion as a continuous, self-directed sequence. The significance here is not technical novelty, since the underlying methods are well-worn, but architectural: an AI agent replaced the human element that ransomware has always required.

The environments targeted shared a recognizable profile: AI-adjacent infrastructure left internet-facing, legacy services running on default or unchanged credentials, and known vulnerabilities that had gone unaddressed. That profile is not rare. It describes a wide cross-section of organizations that have scaled their technology footprint faster than their security practices. JADEPUFFER found and exploited exactly that gap, and it did so without a specialist operator guiding each step.

The takeaway for stakeholders is straightforward: the conditions that made this attack possible are common, the tooling that enabled it will only become more accessible, and waiting for greater sophistication before responding is not a viable posture.

- Patch known vulnerabilities in internet-facing AI infrastructure immediately; do not leave AI orchestration platforms exposed without current security updates.
- Remove default credentials from all internal services: object storage, configuration platforms, and databases are active targets when left at factory settings.
- Keep AI platforms isolated from production secrets; provider API keys, cloud credentials, and database connection strings stored in AI orchestration environments create a direct path to critical systems.
- Replace default cryptographic keys and signing tokens in configuration management services such as Nacos before exposing them to any network.
- Restrict database administrative interfaces from internet exposure entirely and apply network-layer access controls.
- Move from periodic security reviews to continuous runtime monitoring since agentic attackers operate faster than human-paced assessment cycles can detect.

TIDE Team Analysis

JADEPUFFER is categorized as an Agentic Threat Actor, which is a classification applied when an AI agent, rather than a human or a fixed automated script, serves as the primary driver of offensive activity. The campaign proceeded in two phases against two distinct targets, with the first system compromised specifically as a pathway into the second.

Initial access was established by exploiting CVE-2025-3248, an unauthenticated code execution flaw in the Langflow platform. Langflow is an open-source framework used to build AI-driven applications and agent workflows, and it sits in an attractive position for attackers: deployments routinely carry API keys for major AI providers, cloud platform credentials, and internal service configurations as part of normal operations, often without the network controls or secrets hygiene applied to traditional production systems. Once execution was achieved, the agent performed systematic host profiling and conducted a parallel sweep of the environment for multiple secret categories. These include AI provider keys, credentials for cloud platforms spanning both major Western providers and several Chinese services, database connection strings, and cryptocurrency-related files.

The agent then extracted data directly from Langflow's own database backend, gathering stored credentials and user account information, before probing the internal network for adjacent services. A MinIO object storage instance was reached at its default internal address using the service's default administrative credentials. The agent worked through the available storage buckets methodically, ultimately retrieving configuration and credential files whose names suggested high value. Before moving on, a persistence mechanism was installed via the system scheduler, configuring recurring outbound contact with attacker-controlled infrastructure.

The production server targeted in the second phase ran a MySQL database alongside an Alibaba Nacos instance (a service discovery and configuration management platform with a documented history of authentication weaknesses). The default cryptographic signing key used by Nacos for session tokens has been publicly known for years and remains unchanged in many deployments, allowing an attacker to forge valid administrative credentials without any access to the system. The agent pursued access through several simultaneous methods: leveraging the known authentication bypass vulnerability class, constructing forged tokens using the default key, and inserting a backdoor administrator account directly into the Nacos database through the root-privileged MySQL connection it already held.

An initial attempt to create this backdoor account failed because of a library dependency issue. The agent identified the specific cause, constructed a corrected approach, and had a functioning account within roughly half a minute with no human guidance nor external intervention. Once administrative access was confirmed, the agent encrypted the full set of Nacos service configuration records using a database-native encryption function, removed the original data tables along with historical records, and created an extortion notice inside the database containing a Bitcoin payment address and contact information. The encryption key used was randomly generated, appeared once in the agent's execution output, and was discarded. This in turn left the victim with no path to recovery even if a ransom was to be paid. The agent then proceeded to drop entire database schemas, selecting targets based on apparent data value.

Several behavioral characteristics collectively support the assessment that this operation was driven by an autonomous AI model rather than human direction or scripted tooling. Throughout the operation, the agent produced code annotated with natural-language explanations, such as describing the reasoning behind target selection, estimating the relative value of different databases, and narrating what each step was intended to accomplish. This

kind of self-commentary appears naturally in AI-generated code and is essentially absent from human-written attack tooling. More telling still was the pattern of error handling: each time a step failed, the agent's next action corrected the specific underlying cause rather than issuing a generic retry. A database drop blocked by a referential integrity constraint was followed immediately by a payload that suspended that constraint, completed the drop, and restored the original setting. The total operation encompassed several hundred distinct payloads executed within a condensed window, with a coherence and adaptive quality inconsistent with either a human operator or a prewritten script.

Why It Matters

The practical barrier to running a ransomware operation has historically been human expertise. Chaining together exploitation, internal navigation, persistence, and data destruction required either a skilled operator or a team that could cover those disciplines collectively. JADEPUFFER removes that requirement. An AI agent can now work through each of those phases autonomously, meaning an operator no longer needs to understand what they are doing at any individual step since they only need to configure the agent and point it at a target. When the compute cost of running that agent is covered by stolen credentials, the financial barrier disappears alongside the technical one.

The victim profile in this campaign carries a direct implication for organizations assessing their own exposure. The systems compromised were not obscure or unusually difficult to secure. Instead, they were internet-facing AI infrastructure with a known unpatched vulnerability, a configuration service still using a default signing key that has been publicly documented for years, and internal storage accessible with credentials that had never been changed from factory defaults. These conditions exist broadly across organizations that have prioritized deployment speed over security rigor, particularly in the AI and cloud-native infrastructure space where tooling is adopted quickly and hardening practices often lag. The shift to agentic attackers means that the historical assumption that neglected but obscure systems are unlikely to be targeted no longer holds. An agent can work through the full catalog of known vulnerabilities across all exposed systems at negligible cost per attempt.

AI infrastructure itself has become a primary attack surface that many organizations have not yet incorporated into their formal threat models. Platforms used to build and orchestrate AI applications tend to accumulate sensitive credentials as a byproduct of their function, such as keys for commercial AI services, access credentials for cloud environments, database connection strings, and infrastructure configuration data. A single compromised node of this type can give an attacker everything needed to reach production systems, as JADEPUFFER demonstrated by moving from a Langflow server to a production database without acquiring any additional access through external means.

The temporal dimension of this threat also warrants attention. Security programs built around human-paced review cycles (e.g., monthly, quarterly) are structurally mismatched to an attacker that can complete an entire extortion cycle within a single session and correct failed steps in under a minute. Criminal actors face no procurement delays, compliance reviews, or organizational approval processes when adopting new tooling. As agentic offensive capabilities become more packaged and accessible, adoption across the threat actor spectrum will be rapid, and the window between a new capability emerging and it being used at scale against common infrastructure will compress further.

How to Respond

- Patch CVE-2025-3248 on all Langflow instances immediately and remove code execution or validation endpoints from internet exposure. Extend the audit to all other known Langflow CVEs and similar AI orchestration platforms in the environment.
- Audit and relocate secrets stored in AI platform environments. Provider API keys, cloud credentials, and database connection strings should be moved to a dedicated secrets manager scoped to least privilege, which means no internet-facing service should carry standing administrative credentials.
- Harden all Nacos deployments: replace the default JWT signing key, restrict the service to internal networks only, remove root-level database access, and apply patches for CVE-2021-29441 and related authentication bypass variants.
- Conduct a default credential review across all internal services: object storage, configuration platforms, message brokers, and supporting databases. Remediate any unchanged factory defaults found.
- Restrict database administrative ports to internal networks only, enforce strong unique credentials, and apply source IP controls on all management interfaces.
- Implement egress filtering on application hosts to prevent compromised servers from making arbitrary outbound connections or reaching external staging infrastructure.
- Deploy runtime behavioral monitoring capable of alerting on: unexpected outbound connections from application hosts, schedulers executing network calls, bulk table deletions, and schema-level database changes.

What UltraViolet Cyber is Doing

- Tracking new CVEs and high impact vulnerabilities, analyzing and deploying public Proof-Of-Concept code against custom built targets.
- 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