ZEROFOX INTELLIGENCE

Brief

JADEPUFFER and the Arrival of Agentic Ransomware Threats

B-2026-08-10a

Classification: TLP:CLEAR

Criticality: LOW

Intelligence Requirements: Threat Actor, Dark Web, Artificial Intelligence

August 10, 2026

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were identified prior to 10:00 AM (EDT) on August 7, 2026; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

Brief | JADEPUFFER and the Arrival of Agentic Ransomware Threats

Executive Summary

Threat actors have almost certainly crossed a ransomware tradecraft threshold with the first documented end-to-end operation driven by a large language model (LLM) agent. This marks the beginning of a capability shift that is very likely to reshape the affiliate tier of the ransomware ecosystem within the next six to 12 months. While ZeroFox and peer research teams have tracked an explosion of weaponized LLM chatbots and stolen artificial intelligence (AI) credential trade throughout the first half of 2026 on dark web forums and Telegram channels, we have not yet observed packaged agentic-attack toolkits offered for sale. Currently, the payloads autonomous agents produce are unreliable and likely to generate operationally broken extortion demands. ZeroFox predicts that the next attributed agentic ransomware campaign will likely originate from a previously unknown or low-reputation threat actor cluster rather than from a recognized top-10 operator.

Details

On July 1, 2026, cloud security firm Sysdig disclosed that a likely financially motivated threat actor known as “JADEPUFFER” used an autonomous AI agent to conduct reconnaissance, credential theft, lateral movement, persistence, and destructive

database extortion against a single unnamed victim.[1] The threat actor gained initial access through a widely known and long-patched vulnerability, underscoring that the novelty of the operation lies in its orchestration rather than in any single technical component. The entry point was CVE-2025-3248, an unauthenticated remote-code execution (RCE) flaw in the open-source Langflow framework that was patched by the vendor in April 2025 and added to the Cybersecurity and Infrastructure Security Agency's Known Exploited Vulnerabilities catalog in May 2025.[2]

- Sysdig reports the operator delivered a series of over 600 discrete payloads through the vulnerable endpoint, with each payload generated dynamically by an LLM rather than drawn from a fixed toolkit.[3]

The operator then likely pivoted from the compromised Langflow host to a separate internet-facing production server hosting a MySQL database and an Alibaba Nacos configuration service, using root credentials of unknown provenance.[4] The pivot itself is significant, as it establishes that the threat actor was likely not restricted to a single opportunistic foothold and possessed at least some prior intelligence about the target environment.

- Sysdig stated it could not determine how the operator obtained the MySQL root credentials and has separately confirmed to industry press that a human—not the LLM agent—set up the operation, chose the victim, and provisioned the command-and-control and staging infrastructure.[5]

The intrusion chained an unauthenticated RCE flaw in the AI development framework Langflow to a production configuration database; ultimately, the destructive phase concluded with the AI agent encrypting all 1,342 Nacos service configuration items using MySQL's built-in Advanced Encryption Standard (AES) encryption function, dropping the original configuration and history tables, and creating a ransom note table containing a

[1] hXXps://www.bleepingcomputer[.]com/news/security/jadepuffer-ransomware-used-ai-agent-to-automate-entire-attack/
[2] hXXps://www.sysdig[.]com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion
[3] hXXps://www.csoonline[.]com/article/4193195/this-ai-agent-autonomously-hacked-a-network-adapted-on-the-fly-and-demanded-a-ransom.html
[4] hXXps://www.sysdig[.]com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion
[5] hXXps://cyberscoop[.]com/sysdig-judepuffer-ai-agentic-ransomware-attack/

Bitcoin payment address and a Proton Mail contact. The encryption key was generated randomly, printed once to standard output, and never persisted or transmitted anywhere the operator could recover it[6]—meaning that providing a working decryption key was almost certainly impossible, regardless of whether the victim paid the demanded ransom.

- The Bitcoin address included in the ransom note is a well-known example address that appears throughout Bitcoin's own developer documentation, an anomaly Sysdig suggested is consistent with either LLM hallucination from training data or a deliberate operator choice to use a sweep wallet resembling a public example.[7]

On July 20, 2026, Sysdig disclosed that the actors returned to the same compromised Langflow instance approximately three weeks after the initial intrusion and deployed a purpose-built ransomware family specifically engineered to encrypt trained AI model artifacts, training datasets, and vector databases.[8] The operators staged a Go-based ransomware binary (tracked as EncForge) that Sysdig characterizes as built specifically for AI and machine learning infrastructure rather than adapted from a general-purpose encryptor[9]—an evolution ZeroFox assesses transforms JADEPUFFER from a proof-of-concept into an active campaign

- The re-entry itself is significant because it demonstrates the operator almost certainly retained access to, or was able to re-establish access to, the original victim environment despite public disclosure of the initial intrusion.
- EncForge is purpose-built to encrypt AI model artifacts, weights, vector databases, and training datasets across all major machine learning frameworks, with Sysdig estimating damages of USD 75,000 to USD 500,000 per affected model and potential losses of weeks or months of training work per victim.[10]

[6] hXXps://www.infosecurity-magazine[.]com/news/researchers-first-agentic/
[7] hXXps://www.sysdig[.]com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion
[8] hXXps://www.bleepingcomputer[.]com/news/security/jadepuffer-agentic-attacks-now-target-ai-model-data-with-ransomware/
[9] hXXps://www.sysdig[.]com/blog/jadepuffer-evolves-the-agentic-threat-actor-deploys-ransomware-built-to-destroy-ai-models
[10] *Ibid.*

- EncForge corrected the unrecoverable-key error from the first campaign and adopted encryption and ransom-note conventions consistent with established ransomware families.[11] ZeroFox assesses this shift reflects deliberate operator learning between engagements and very likely moves JADEPUFFER's tradecraft closer to that of professional ransomware operations.

The operational significance is very likely not the individual techniques, which remain unremarkable, but the threat actors' demonstrated ability to iterate autonomous attack tooling on a three-week development cycle and to specifically target the AI infrastructure ecosystem for extortion. Threat actors have historically required either deep in-house expertise or access to human affiliates capable of chaining reconnaissance, exploitation, credential handling, lateral movement, and extortion into a single coherent operation; this expertise gap has very likely been the primary throttle on ransomware supply for years.

- JADEPUFFER demonstrates that this gap can now be closed by an LLM agent operating on stolen or rented access to commercial AI models. Further, the EncForge tool release demonstrates that agentic threat actors can build durable, purpose-designed malware families rather than only relying on generated one-off payloads.

ZeroFox assesses that criminal marketplace economics will dictate how quickly the capability demonstrated by the JADEPUFFER incidents spreads. ZeroFox and peer research teams have tracked an explosion of weaponized LLM chatbots and stolen AI credential trade throughout the first half of 2026 on dark web forums and Telegram channels, but we have not yet observed packaged agentic-attack toolkits offered for sale. This gap between capability demonstration and marketplace commoditization is almost certainly the defining strategic window for defenders, and it is very likely closing.

[11] hXXps://www.sysdig[.]com/blog/jadepuffer-evolves-the-agentic-threat-actor-deploys-ransomware-built-to-destroy-ai-models

Observations on Dark Web Marketplaces

Dark web forum and Telegram channel threat actors currently sell the ingredients for agentic attacks but do not yet offer the finished product, a distinction ZeroFox assesses is significant when evaluating the risk of the JADEPUFFER incidents. Coverage of the incidents has driven a wave of vendor commentary framing agentic ransomware as an imminent commodity threat, but underground marketplace inventory as of Q3 2026 does not support that assessment. Rather, ZeroFox has observed that threat actors are actually purchasing (in volume) jailbroken conversational LLMs and stolen access to legitimate model providers—neither of which is equivalent to a turnkey autonomous ransomware toolkit.

Weaponized LLM chatbots have been among the fastest-growing category of criminal AI tooling for the past two years and continue to dominate marketplace inventory, with the WormGPT brand family, FraudGPT successors, and 2025–2026 entrants such as KawaiiGPT and WormGPT 4 offered at subscription prices between approximately USD 50 per month and USD 220 for lifetime access.[12] Former FBI Cyber Deputy Director Cynthia Kaiser reported on Halcyon research at Infosecurity Europe in June 2026 that tracked mentions of AI on dark web forums; she noted such chatter rose from 38 posts in December 2025 to 1,486 in February 2026, an increase drawn from approximately 4,000 entries across 77 Telegram channels, 20 dark web forums, and five underground markets.[13]

- These tools generate phishing lures, malicious code, and social engineering scripts on demand, but they are not autonomous agents and do not chain intrusion phases.

The more consequential trend signified by the JADEPUFFER incidents is almost certainly the maturation of the market for stolen access to commercial LLM providers, a technique Sysdig originally coined as "LLMjacking" in May 2024.[14] Threat actors almost certainly now routinely harvest OpenAI, Anthropic, Google Gemini, and DeepSeek API keys from

[12] hXXps://www.theregister[.]com/security/2025/11/25/lifetime-access-to-wormgpt-4-costs-just-220/2507953
[13] hXXps://www.infosecurity-magazine[.]com/news/cybercrime-ai-tools-surge-3800/
[14] hXXps://www.sysdig[.]com/learn-cloud-native/what-is-llmjacking

misconfigured cloud environments and exposed AI development frameworks, reselling them at low unit prices through dedicated marketplaces and reverse-proxy services.

Security researchers have observed threat actors running AI-orchestrated malware development operations that pre-date the JADEPUFFER disclosure, indicating that agentic tradecraft is emerging along multiple parallel tracks rather than via a single lineage. On June 2, 2026, Sophos disclosed that a separate threat actor had operationalized an AI-driven framework to develop and refine ransomware payloads capable of evading major endpoint detection products.[15]

- The analytical distinction is almost certainly that the Sophos framework represents AI-orchestrated development with human execution, while JADEPUFFER represents AI-orchestrated execution with human setup; these two hybrid patterns are likely eroding the "human at the keyboard" ransomware model from opposite ends.

However, ZeroFox has not yet observed a packaged, agentic ransomware kit for sale or an "autonomous pentest agent" advertised as a service on any dark web venue to which we have access. ZeroFox assesses that the absence of such a product is a genuine finding rather than a coverage gap and likely reflects that the operators currently capable of building these tools are keeping them private for direct use.

- Reporting that names established ransomware brands as having "confirmed AI agent integration" traces to secondary marketing content and does not withstand primary-source scrutiny; ZeroFox flags this as a specific mis-framing clients are likely to encounter and should discount.

From a defensive perspective, current underground activity suggests that threat actors likely remain in an accumulation phase rather than an operational maturity phase. However, the individual building blocks of a JADEPUFFER-class capability—including jailbroken or stolen AI models, exposed AI development frameworks, and exploit-chaining techniques—are increasingly obtainable through separate channels. Although researchers have not yet observed these capabilities consistently packaged into a scalable criminal service, recent JADEPUFFER activity and the appearance of

[15] hXXps://www.helpnetsecurity[.]com/2026/06/02/ai-agents-edr-evasion-techniques/

EncForge indicate that the barriers to such integration are likely decreasing. Defenders should therefore expect continued experimentation and a gradual transition toward more operationalized AI-enabled intrusion tooling.

Threat Actor Capabilities Expanding

ZeroFox has observed that threat actors adopting new tradecraft historically follows a diffusion curve that begins with a single operator, moves through a validation phase of one to three imitators, and reaches broad affiliate-tier adoption between roughly 12 to 18 months. JADEPUFFER is likely the beginning point of that curve rather than the mid-point.

- The most useful analogue is very likely double-extortion ransomware, which the Maze group pioneered in late 2019 and had been adopted by at least 15 ransomware families operating across roughly 1,200 incidents by the end of 2020.[16]
- The ransomware-as-a-service (RaaS) model itself followed a slower three-year diffusion from initial appearance in 2016 to majority-market operation by 2019; structurally, it almost certainly had the same outcome, as it lowered the skill floor for entry.

The specific capability JADEPUFFER demonstrates—chaining reconnaissance, credential theft, lateral movement, persistence, privilege escalation, and destruction without deep operator expertise in any individual step—almost certainly targets the exact bottleneck that has historically constrained ransomware supply. Threat actors operating at the affiliate tier of RaaS programs are typically competent in one or two intrusion phases but rely on either playbooks or in-group support to complete the full kill chain, and this tactical execution gap has almost certainly been the primary reason low-skilled operators either fail to convert access into extortion or abandon operations mid-intrusion.

- An LLM agent that can compose the intrusion chain live using general knowledge about a class of applications it has previously trained on almost certainly removes that gap without requiring the operator to develop the underlying expertise.

[16] hXXps://www.zscaler[.]com/resources/security-terms-glossary/what-is-double-extortion-ransomware

Threat actors iterating agentic tooling between successive campaigns represents a distinct and very likely more concerning capability tier than any single autonomous operation, and the JADEPUFFER return demonstrates this iteration cycle is measured in weeks rather than months. The three-week interval between the initial JADEPUFFER disclosure and the EncForge follow-on is shorter than the typical development cycle for a new ransomware family and likely suggests the operator either possessed the EncForge tooling prior to the first campaign or was able to develop it rapidly using the same agentic infrastructure that supported the initial intrusion.

Prior AI-assisted ransomware operations have illustrated both the potential and the current limits of this capability. In August 2025, Anthropic disclosed that a threat actor tracked as "GTG-2002" used its Claude Code product to conduct data-extortion operations against at least 17 organizations; ransom demands ranged from USD 75,000 to over USD 500,000, with a human operator directing the campaign throughout.[17] In November 2025, Anthropic further disclosed that a Chinese state-linked threat actor tracked as "GTG-1002" conducted espionage against approximately 30 targets, with Claude executing an estimated 80–90 percent of tactical work autonomously—marking the highest publicly attributed autonomy level prior to JADEPUFFER.[18]

- ZeroFox assesses these two disclosures bracket the current capability envelope, with JADEPUFFER positioned closer to the autonomous end and the criminal ecosystem as a whole sitting closer to the human-steered end.

The affiliate-tier operators most likely to adopt agentic tooling first are those already comfortable operating on stolen or rented infrastructure and unconstrained by the operational security discipline that established ransomware brands enforce; ZeroFox has tracked several 2025–2026 cases that support this profile.

- Check Point Research has publicly assessed that the FunkSec ransomware group, which surfaced in late 2024 and claimed over 85 victims in December 2024 alone, used AI-assisted code development for its Rust-based ransomware and associated tooling.[19]

[17] hXXps://www.darkreading[.]com/cyberattacks-data-breaches/anthropic-ai-automate-data-extortion-campaign
[18] hXXps://www.anthropic[.]com/news/disrupting-AI-espionage
[19] hXXps://research.checkpoint[.]com/2025/funksec-alleged-top-ransomware-group-powered-by-ai/

- FunkSec's operational profile—inexperienced operators, recycled hacktivist data leaks, unusually low ransom demands averaging approximately US 10,000, and sloppy key handling that enabled Avast to release a public decryptor[20]—is consistent with the profile ZeroFox expects for early agentic-ransomware adopters.
- In a 2026 case, Sicarii ransomware was reportedly shipped with an AI-introduced encryption bug that made victim recovery impossible even with a working decryptor[21], a failure mode that also appeared in JADEPUFFER's first campaign key-handling but was corrected in the EncForge follow-on.

Established, top-tier RaaS operators are unlikely to be first movers on fully autonomous operations for reasons of business logic rather than capability. Groups such as Qilin, Akira, and LockBit derive substantial revenue from a functioning affiliate program and negotiated ransom payments, and both revenue streams almost certainly depend on the victim believing that payment will produce a working decryptor.

Large, established ransomware collectives are almost certainly integrating AI and agentic automation into discrete segments of the attack chain rather than pursuing a JADEPUFFER-style, end-to-end autonomous operation as an initial adoption model. This phased approach is consistent with the operational conservatism these groups have historically shown toward any technique that could jeopardize the negotiation and payment infrastructure on which their revenue model depends. The endpoint detection and response (EDR)-evasion framework identified by Sophos exemplifies the ongoing adoption of segmented AI-orchestrated tradecraft. In this instance, threat actors utilized coordinated LLM agents to expedite the iterative development and validation of malicious payloads, yet maintained a human presence to oversee deployment logistics and victim negotiations. This hybrid model demonstrates AI functioning as a tactical accelerator for specific kill-chain phases rather than as a substitute for human-directed operational control.

- Reputable, established collectives are very likely to avoid publicly advertising AI or agentic tooling in their operations—discretion that is consistent with the

[20] hXXps://www.gendigital[.]com/blog/insights/research/funksec-ai
[21] hXXps://www.halcyon[.]ai/ransomware-alerts/alert-sicarii-ransomware-encryption-key-handling-defect

operational security discipline that separates top-tier RaaS brands from opportunistic or hacktivist-adjacent actors.

- This creates a durable observability gap: dark web and Telegram monitoring functions well as a lens on the retail market for jailbroken models and stolen credentials but is currently poor at detecting or observing private, in-house agentic tooling that a well-resourced group has no commercial incentive to expose. ZeroFox assesses that the absence of established-actor chatter about agentic capability should not be read as evidence it is not occurring.

ZeroFox assesses with moderate confidence that near-term agentic ransomware development will bifurcate along two tracks distinguished more by operational discipline than technical sophistication. Low-credibility and inexperienced actors are likely to continue publicly testing JADEPUFFER-style, fully autonomous attack chains, generating further attributable incidents that are individually unreliable (broken encryption, hallucinated payment details, or premature exposure) and disproportionately likely to be identified and disclosed by researchers because these actors lack the tradecraft to keep experimentation quiet. Concurrently, established and well-resourced collectives are very likely to develop agentic capability privately, deferring deployment until the tooling reaches a reliability threshold consistent with their existing negotiation and payment operations.

- This produces a paradoxical near-term dynamic in which the publicly visible cases of agentic ransomware—JADEPUFFER included—likely minimize the true state of development, since the most capable operators have the strongest incentive to keep their agentic tooling out of view.

Expansion Outlook

Threat actors are almost certain to expand the JADEPUFFER attack pattern beyond Langflow to other internet-exposed AI development and orchestration platforms within the next two quarters, and the aperture of viable initial access points is very likely wider than most clients currently model. The AI-adjacent internet-facing attack surface is almost certainly large, growing, and unpatched at scale.

Threat actors targeting AI model artifacts as a distinct asset class represents a specific evolution in ransomware target selection that ZeroFox assesses is likely to be adopted beyond JADEPUFFER within the next two quarters. Trained models, fine-tuned adapters, vector databases, and curated training datasets represent asset classes that are often poorly backed up, difficult to reconstruct, and highly valuable to the victim organization—a combination that matches the profile ransomware operators have historically preferred.

Threat actor adoption of the JADEPUFFER pattern is likely to be shaped over the next six to 12 months by three specific developments ZeroFox is tracking as leading indicators, with two of the three showing partial early activation at the time of this writing.

- The first is the appearance of a packaged agentic-attack framework or "autonomous pentest agent" advertised on a reputable underground forum, which would mark the transition from private tooling to marketplace commodity and would likely signal the diffusion inflection point. ZeroFox has not yet observed this indicator.
- The second is a sustained decrease in the underground price of stolen LLM API access, which would compress the operating cost of agentic attacks toward zero and remove the economic barrier for opportunistic operators. This indicator has shown partial activation, as ZeroFox has observed LLMjacking marketplace maturation through mid-2026.
- The third is the emergence of subsequent attributable agentic ransomware operations (whether against a new victim or a member sector) by other distinct actors that would mark the transition from an isolated operator's tradecraft to a diffused pattern expanding across the landscape. This indicator remains largely unactivated as JADEPUFFER escalated its own campaign, rather than the operation being replicated by another actor.

Threat actors are unlikely to abandon traditional ransomware tradecraft in favor of agentic operations in the near term, and it is likely that the two modes will coexist rather than agentic operations acting as a tradecraft substitute or replacement. Traditional operations almost certainly remain more profitable per successful intrusion, produce

more reliable extortion outcomes, and provide more established negotiation and payment infrastructure than agentic operations are capable of to date.

The more accurate framing is that agentic tooling will likely expand the total pool of viable ransomware operators by lowering the entry barrier rather than displacing existing operators, serving as a supply-side expansion rather than a substitution. ZeroFox assesses this expansion will likely disproportionately affect organizations with exposed AI development infrastructure, holding valuable trained AI model artifacts, and in sectors already targeted by low-skilled, opportunistic actors.

| Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	**Sources may use** **TLP:RED** when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	**Sources may use** **TLP:AMBER** when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	**Recipients may NOT share** **TLP:RED** with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	**Recipients may ONLY share** **TLP:AMBER** information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. **Note that** **TLP:AMBER+STRICT** restricts sharing to the organization only.

	Green	Clear
WHEN SHOULD IT BE USED?	**Sources may use** **TLP:GREEN** when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	**Sources may use** **TLP:CLEAR** when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	**Recipients may share** **TLP:GREEN** information with peers and partner organizations within their sector or community but not via publicly accessible channels.	**Recipients may share** **TLP:CLEAR** information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%

Appendix C: ZeroFox Intelligence Threat Actor Reputation Scale

Untested	Moderately Credible	Well-regarded	Prominent
Has garnered no reputation; credibility cannot be determined.	Has made up to 10 transactions; has been active on forum for at least three months.	Has at least 10 transactions; has been active on the forum for three months to one year.	One of the most well-known and credible threat actors on the site; long-term, established presence on the forum of more than one year.