

Phantom Squatting: AI-Hallucinated Domains as a Software Supply Chain Vector

 unit42.paloaltonetworks.com/phantom-squatting-hallucinated-web-domains

Keerthiraj Nagaraj, Diva-Oriane Marty, Beliz Kaleli, Oleksii Starov

June 30, 2026

Executive Summary

Unit 42 researchers found that large language models (LLMs) consistently hallucinate web domains for legitimate brands. Adversaries are actively weaponizing this vector by registering these nonexistent domains to intercept traffic generated by AI systems. We call this phenomenon phantom squatting, and it poses a significant risk to the software supply chain.

Our proactive monitoring of registration for high-priority hallucinated domains yielded real-world detections across multiple sectors. We were able to predict use of these domains from 18–51 days ahead of adversary registration.

A standout case reveals an attacker who leveraged an AI coding assistant to build a full phishing kit named Montana Empire. This kit targeted a domain our detection pipeline identified as a high-risk hallucination target 23 days earlier, demonstrating the full cycle from AI-assisted attack development to LLM-hallucinated domain prediction.

To detect the risk posed by phantom squatting, we analyzed 913 global brands and executed 685,339 URL queries across multiple configurations of two distinct LLM models. This generated 2.1 million URLs and revealed over 13,229 confirmed malicious URLs. Furthermore, we discovered approximately 250,000 hallucinated domains that remain unregistered, presenting a significant opportunity for adversaries to exploit the software supply chain through preemptive registration.

Palo Alto Networks customers are better protected from phantom squatting through the following products and services:

- [Advanced WildFire](#)
- [Advanced URL Filtering](#) and [Advanced DNS Security](#)
- [Prisma AIRS](#)
- [Koi Agentic Endpoint Security](#)

The [Unit 42 AI Security Assessment](#) can help empower safe AI use and development.

If you think you might have been compromised or have an urgent matter, contact the [Unit 42 Incident Response team](#).

Introduction: LLMs as Supply Chain Dependencies

The Expanding AI Trust Surface

The software supply chain threat landscape is shifting. For decades, supply chain attacks focused on predictable artifacts such as tampered build tools, malicious dependencies and compromised update servers. Defenders built protections around these predictable attack surfaces using package integrity checks, signed binaries and dependency auditing tools.

However, this model is becoming less effective. LLMs are no longer peripheral utilities, they are active participants in the software development lifecycle.

People consult AI coding assistants for documentation links. In doing so AI agents perform autonomous web research on behalf of developers, then formulate and execute HTTP requests against URLs the models themselves generate.

Enterprise continuous integration and continuous delivery (CI/CD) pipelines integrate AI assistants that recommend third-party service endpoints. For example, a developer querying a pipeline assistant to configure a cloud deployment notification might receive a recommended webhook URL such as `hxxps[:]//api.build-notifier[.]io/v1/pipeline/events`. Such a URL could be entirely fictitious and an adversary could have pre-registered it to intercept automated build telemetry or secrets.

In each case, downstream consumers often trust the LLM's output including the URLs it generates, without independent verification. This situation fundamentally alters the attack surface. When an LLM produces a URL, that artifact may be:

- Ingested directly by autonomous AI agents that retrieve the resource
- Integrated by developers into production-grade code
- Suggested by AI coding assistants as the authoritative endpoint for third-party services
- Included in documentation generated through large-scale automation

In these scenarios, an LLM functions as a trusted supply chain dependency. However, as with any trusted architectural component, it is susceptible to systematic exploitation.

From Slopsquatting to Phantom Squatting: Extending the AI Supply Chain Attack Taxonomy

Prior [research on slopsquatting](#) established the foundational attack pattern. LLMs frequently hallucinate software package names that do not exist in any legitimate registry.

Phantom squatting extends this adversarial logic from software packages to web infrastructure. Just as an LLM might hallucinate a library name, it can generate fictitious domains for web portals, API endpoints or corporate services for a target brand.

Throughout this article, we use the term phantom domain to specifically refer to a hallucinated domain that an adversary has or could weaponize.

The adversarial logic is illustrated by the following scenarios:

- A coding assistant generates a plausible but unregistered benefits portal URL, allowing an adversary to preemptively register it.
- An AI research agent produces a plausible banking portal domain that an adversary could have already registered to capture traffic.
- A developer integrates an AI-generated API endpoint into their code, unknowingly directing application data to an attacker-controlled server.

This is no longer a theoretical risk. Our research confirms this vector is currently active in the wild.

Why Existing Supply Chain Defenses Miss This Threat

Typically, URL filtering and threat intelligence frameworks operate under a critical, shared assumption, that malicious infrastructure possesses a detectable reputation. Typical block lists rely on historical reports of malicious activity, while threat feeds require a domain to be observed within an active campaign before classification. Reputation scoring models require a domain to maintain a presence long enough to accumulate telemetry signals.

A phantom domain effectively exploits a zero-reputation bypass. At the moment an adversary registers and weaponizes a hallucinated domain it:

- Carries no threat intelligence history
- Has not established a reputation score
- Lacks any blocklist entries

The infrastructure is nascent, the content is original and conventional defensive perimeters have no actionable signal. By the time threat intelligence systems synchronize, people have

already been funneled to the site by an AI system they consider authoritative.

This shows the structural advantage of phantom squatting over legacy phishing. The fake domain is born clean because it comes from the LLM's own internal vocabulary. These are the same language patterns that make the model's output seem legitimate.

Threat Model: The Phantom Squatting Attack Lifecycle

Figure 1 shows the phantom squatting attack lifecycle operates across four distinct phases:

- Discover
- Act
- Lure
- Bypass

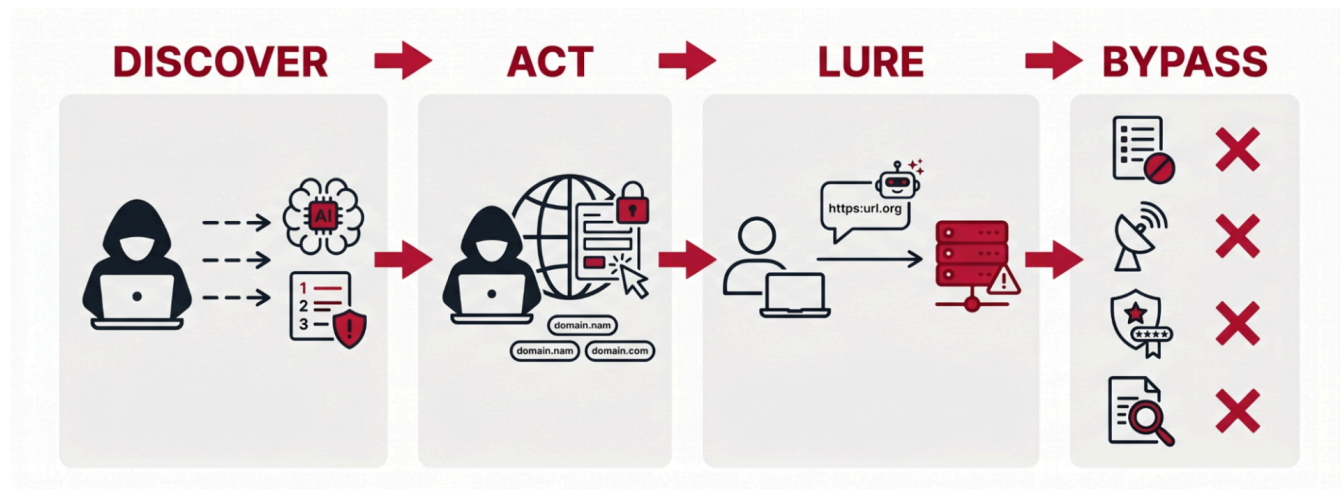


Figure 1. The phantom squatting attack lifecycle across four phases.

Discover: Adversarial Probing of LLM Hallucination Patterns

The adversarial lifecycle begins by mapping a target brand's hallucination surface — the collection of phantom domains an LLM generates in response to diverse prompt strategies. This phase, which we define as adversarial hallucination probing, involves systematically querying models. Attackers could use realistic prompts that mimic everyday user operations, with the primary objective of observing and mapping the resulting hallucination patterns.

Act: Registering Hallucinated Phantom Domains Before Defenders React

Armed with a prioritized inventory of phantom domains, adversaries proceed to preemptively register those most valuable for attacks. For generic top-level domains (TLDs), the barriers to entry are negligible. Registration is both economical and nearly instantaneous. Our analysis

confirms that threat actors operate with significant speed, often well within the window of any feasible defensive response.

In observed real-world telemetry, these domains transitioned from initial registration to active malicious content deployment within hours. In the case of Montana Empire, the adversary had even staged the server-side phishing kit prior to the domain's registration, demonstrating a highly optimized zero-reputation bypass strategy.

Lure: LLMs as Unwitting Attack Delivery Mechanisms

Following the registration and subsequent weaponization of a phantom domain, the LLM itself functions as the primary attack delivery mechanism. Any user or autonomous AI agent that issues a query triggering the hallucinated URL receives an authoritative, high-confidence recommendation to navigate directly to attacker-controlled infrastructure.

This represents a defining characteristic of the phantom squatting threat. The delivery vector bypasses traditional phishing emails, malvertising or [watering hole](#) attacks. Instead, the delivery mechanism is the trusted AI assistant already integrated into the user's workflow.

Consider a scenario where an employee queries for a third-party service endpoint from an AI coding assistant. If the LLM provides a fictitious domain like evilphishing[.]com/auth/login, the exploitation occurs without a single traditional phishing lure. The victim is compromised simply by following a confident recommendation from a system their organization has already formally sanctioned.

Bypass: Zero-Reputation Evasion of Reputation-Based URL Defenses

The final phase of the attack lifecycle relies on a newly registered phantom domain's zero-reputation status, circumventing most conventional URL defenses. As noted earlier, at the moment of registration and initial weaponization, the domain lacks any blocklist entries, threat intelligence history or established reputation score. It has not yet been reported or classified by people.

From a defensive perspective, the infrastructure is nascent and indistinguishable from any legitimate new domain until it has generated sufficient malicious telemetry to trigger a classification signal. By the time threat intelligence systems synchronize, the exploit has already been delivered to victims who relied on the trusted AI assistant's authoritative recommendation.

This structural advantage for attackers is not merely a transient window of opportunity. Sophisticated attackers can maintain this bypass through active evasion techniques, including

redirect cloaking — serving benign content to automated crawlers while targeting human visitors — and the deployment of CAPTCHA-protected infrastructure.

A Proactive Hallucination Discovery Framework

To quantify and operationalize the phantom squatting threat, we engineered a multi-agent discovery framework. This framework simulates the comprehensive attack lifecycle, from adversarial probing to real-world registration detection. Figure 2 shows the discovery pipeline of this framework.

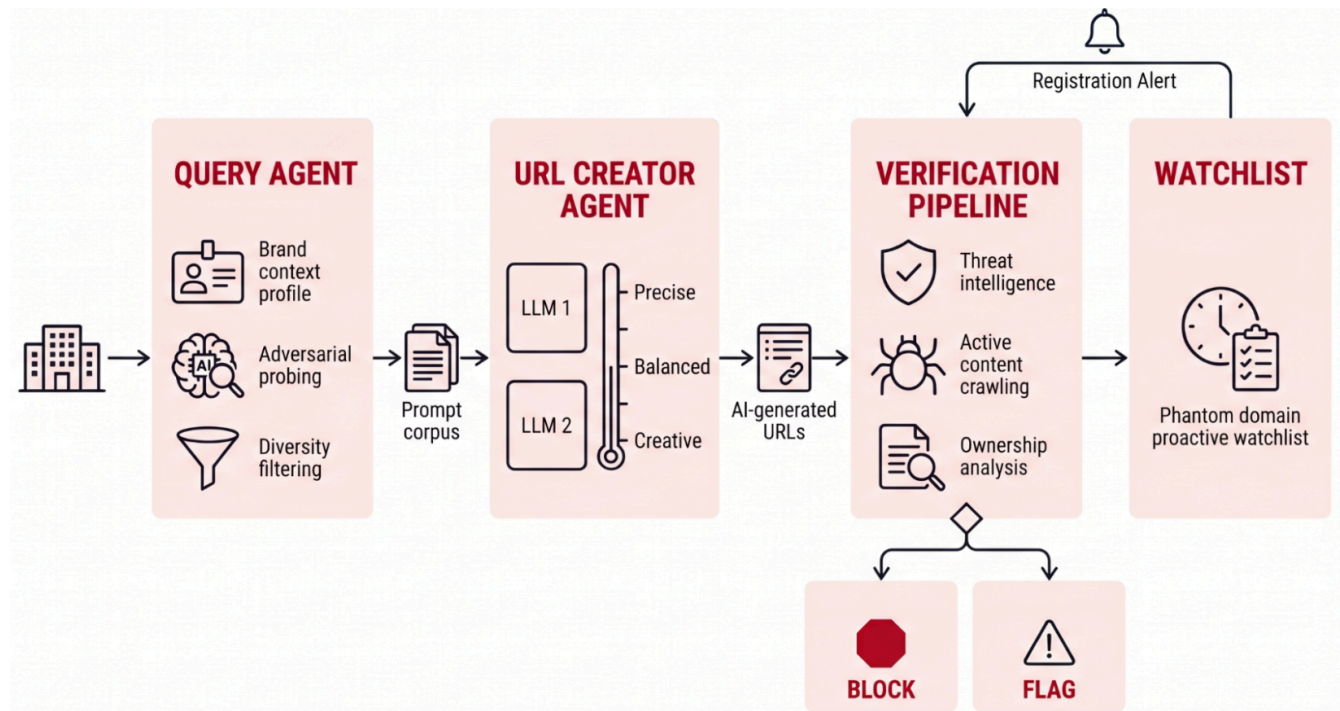


Figure 2. The phantom squatting multi-agent discovery pipeline.

Query Agent: Simulating Attacker Probing

The query agent shown in Figure 2 generates a prompt corpus to probe LLMs. It operates in three main phases.

- **Brand context profile:** The agent researches a brand's products, portals and developer resources. This process ensures prompt references to real services, which helps generate high-fidelity hallucinations.
- **Adversarial probing:** Effective probing requires a diverse set of realistic prompts. Rather than probing randomly, we exploit known LLM failure modes to generate a realistic and diverse set of prompts at scale. These include premise acceptance, authority-framing compliance and the model's tendency to complete narratives with authoritative yet fictitious details.

- **Diversity filtering:** To ensure variety, we use [Jaccard similarity](#) to filter out similar prompts. This broadens the probe of the target's hallucination surface.

This methodology produced 685,339 prompts across 913 global brands.

URL Creator Agent: Mapping Hallucination Behavior Across Models and Temperatures

Prompts from the query agent feed into the URL creator agent. The URL creator agent executes the prompt corpus across multiple LLM providers and a spectrum of [LLM temperature](#) configurations. Our methodology used two distinct LLM families:

- **LLM1:** A production-optimized, mini-class variant of an enterprise LLM from a major technology provider (released April 2025), engineered for high-volume, cost-efficient deployment.
- **LLM2:** A low-latency, lite-class variant of a frontier LLM from a leading AI provider (released June 2025), designed for cost-efficient deployment at scale.

We designate these models as LLM1 and LLM2 throughout this analysis. This distinction is important because both models were released before the malicious domains identified in this research were registered. This confirms that the phantom domains were generated by the models' internal language patterns, not learned from training data. We tested each prompt using three [temperature settings](#) (designated below as T) to test the AI responses:

- **Precise (T = 0.1):** The model is highly predictable, almost always choosing the most likely next token, resulting in consistent and repetitive answers.
- **Balanced (T = 0.7):** This setting mixes predictability with some variability, balancing consistency with a touch of novelty.
- **Creative (T = 1.5):** The model selects from a wider range of less likely words, leading to more imaginative and diverse outputs.

We collected all the URLs found in the LLM responses. If the model didn't provide a URL or said it didn't know the answer, we ignored that specific response. This phase ends with a prioritized list of hallucinated domains that we discovered. The value of these domains to an attacker is determined by two main features:

- **Thermal hallucination persistence (THP):** This measures how consistently the AI generates the same domain name. Domains that appear even when the AI is set to be very precise are high-value targets. This is because the AI is more likely to show these to real users as if they were facts.

- **Cross-model hallucination consensus:** This occurs when different types of AI models all generate the same fictitious domain for the same prompt. If several different models all agree on the same wrong information, it makes that fake domain a much more predictable target for attackers to use.

Verification Pipeline: Multi-Signal Risk Classification

URLs generated by the URL creator agent feed into the verification agent, which assesses multi-signal risk and processes each unique AI-generated URL through an enrichment pipeline that integrates:

- **Threat intelligence:** Category and risk verdicts from threat intelligence systems for existing URLs
- **Active content crawling:** Capturing live page content and screenshots, which are then analyzed by a suite of deep learning models trained to detect malicious signals for existing URLs.
- **Ownership analysis:** Examination of the registrar, registration date, registrant organization, nameservers and privacy status. This data is compared against the legitimate brand's established registration profile.

If a URL exists and exhibits malicious signals, we block it immediately. If a URL shows high-risk indicators, it is flagged for in-line content analysis and added to the proactive watchlist to monitor for changes in registration details or page content. These high-risk indicators include parked pages or insufficient content for a definitive malicious categorization.

We refer to domains not yet registered at the time of analysis as non-existent domains (NXDs). We add these NXDs to a proactive watch list of phantom domains. We then use periodic monitoring of registration event streams to detect when any watchlisted domain is registered.

When a registration event matches a hallucinated phantom domain, an alert is generated and the domain re-enters the verification pipeline for additional analysis. If the newly registered domain proves benign, it is removed from the watch list.

For example, if a legitimate brand registers a domain for defensive purposes or a new product offering, it is considered benign. However, if the ownership or content shows malicious indicators, the domain is assigned a malicious verdict.

Results: Quantifying the LLM Supply Chain Attack Surface for

Phantom Squatting

This section quantifies the phantom squatting attack surface, measured at the domain level rather than the URL level. Although our pipeline extracts millions of unique URLs, the registerable attack surface is at domain level.

Each generated URL undergoes DNS resolution to determine whether it resolves to live infrastructure, NXDs or high-risk endpoints. NXD URLs are then normalized to extract the parent registerable namespace. If that namespace is unregistered, it is enrolled in the phantom domain watchlist.

The subsections below characterize the full risk landscape:

- Confirmed malicious infrastructure served by these models
- The structural composition of the phantom domain inventory
- The model and configuration level factors that govern hallucination volume

Dataset Scale

Our analysis encompasses a dataset of 913 global brands including the following sectors:

- Technology
- Finance
- Healthcare
- E-commerce
- Government
- Gambling
- Logistics

To construct the hallucination corpus, we executed 685,339 adversarial prompts across the LLM1 and LLM2 architectures, yielding 2.1 million unique URLs.

Active Threat Intelligence: Malicious URLs Generated by LLMs

Our discovery pipeline identified that, of the 2.1 million unique URLs produced by the models, threat intelligence systems flagged 13,229 (0.61%) as malicious at the time of analysis.

These results underscore that the risk is not merely theoretical. LLMs are actively recommending known malicious infrastructure to downstream users.

Beyond these confirmed threats, an additional 41,313 URLs (1.90%) were categorized as high risk — including parked domains, adult content and pages with insufficient telemetry —

representing nascent infrastructure or opportunistic targets for adversarial registration.

Figure 3 illustrates the threat landscape of confirmed malicious infrastructure generated by these models.

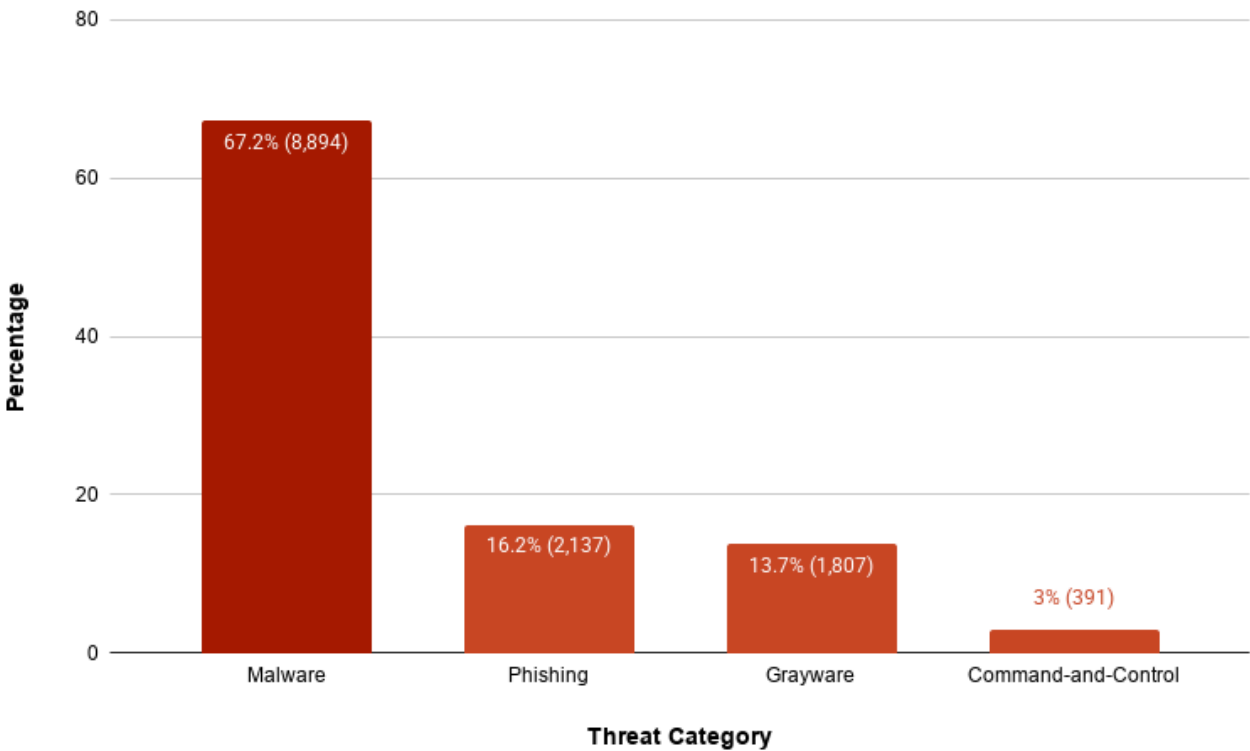


Figure 3. Threat category breakdown of confirmed malicious URLs hallucinated by LLMs.

Malware represents the dominant category at 67.2%, comprising sites used for drive-by downloads, malicious scripts and exploit-kit delivery. Phishing artifacts (16.2%) encompass credential harvesting portals and brand-impersonation sites targeting the global organizations in our analysis. Grayware (13.7%) includes adware distribution and potentially unwanted program (PUP) installers. Of significant concern, command-and-control (C2) infrastructure accounts for 3.0% of identified URLs — a vector of particular risk for autonomous AI agents that may execute web requests to attacker-controlled endpoints when interpreting LLM-generated instructions.

The Phantom Domain Opportunity

Our pipeline revealed that of the 2.1 million unique URLs in our corpus, 809,455 (37.28%) resolve to NXDs — fictitious endpoints generated by LLMs. These 809,455 NXD URLs collapse into approximately 250,000 unique phantom domains after normalization, each representing a discrete, preemptive registration opportunity for an adversary.

The derivation of this dataset is architecturally significant. Approximately 10.8% of NXD URLs (~87,630) constitute pure domain-level hallucinations, where the LLM fabricates an entirely unregistered root namespace. The remaining 89.2% involve subdomain or path-level hallucinations.

To isolate the registerable attack surface, we extracted the parent domain for each artifact. If the parent was unregistered, we enrolled it in our phantom domain watch list. This extraction methodology reduces the 809,455 URL-level NXDs to a tractable inventory of approximately 250,000 registerable phantom domains.

LLM Model Comparison: Hallucination and Threat Rates

Comparative analysis of the two models reveals markedly divergent hallucination profiles despite evaluation against an identical corpus of prompts. LLM1, the production-optimized enterprise model, exhibited a significantly elevated NXD rate of 44.6% across its 1.2 million unique URLs, approximately 17 percentage points above the 27.5% rate observed for LLM2.

Hallucination volume varies substantially by model architecture. However, the confirmed malicious URL rates remained comparable at 0.64% and 0.56%, respectively, indicating that the susceptibility to generating malicious infrastructure is a consistent risk across disparate training lineages.

A consistent pattern emerges regarding high-risk benign URLs, where LLM1 (2.08%) again demonstrates a higher rate than LLM2 (1.67%). This further confirms that LLM1's increased output volume systematically expands the hallucination surface across all risk tiers, extending beyond confirmed malicious infrastructure.

Figure 4 illustrates the comparative landscape of risk across both LLM architectures. It delineates the NXD hallucination rate, the volume of confirmed malicious URLs and the prevalence of high-risk artifacts identified within the corpus.

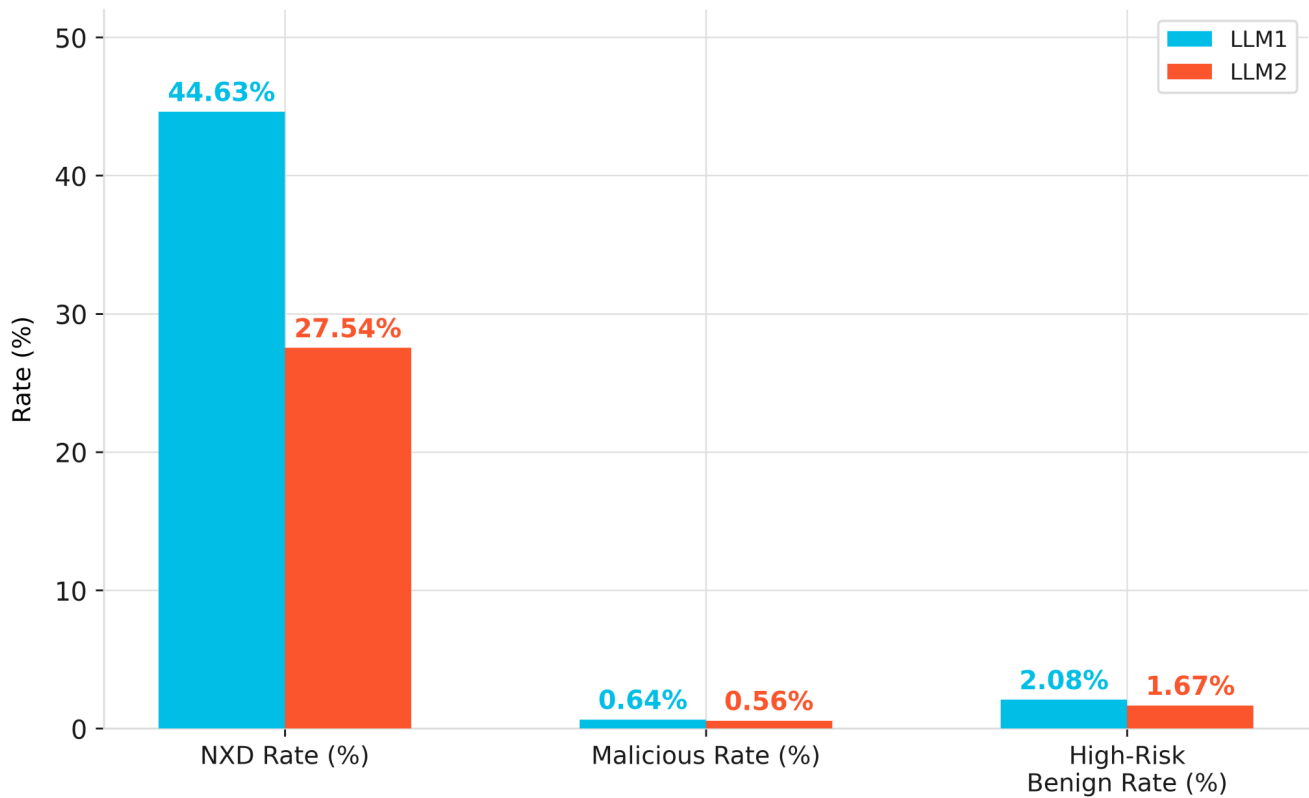


Figure 4. Comparison of LLM results.

Temperature Configuration and Hallucination Risk

LLM inference temperature, the parameter controlling output randomness, quantifiably influences phantom domain generation rates. Across three configuration modes evaluated uniformly, the Creative configuration ($T = 1.5$) yielded a substantially elevated NXD rate of 43.10%, compared to 34.64% for Precise ($T = 0.1$) and 32.52% for Balanced ($T = 0.7$).

Conversely, malicious URL rates remained statistically stable between 0.57–0.63%, suggesting that adversarial content risk is an intrinsic model property rather than a function of entropy. This structural decoupling confirms that while temperature does not drive malicious intent, it significantly amplifies the total hallucination-based exposure surface.

Figure 5 illustrates the impact of inference temperature configuration on both the NXD hallucination rate and malicious URL rate across the three distinct modes used in our discovery pipeline.

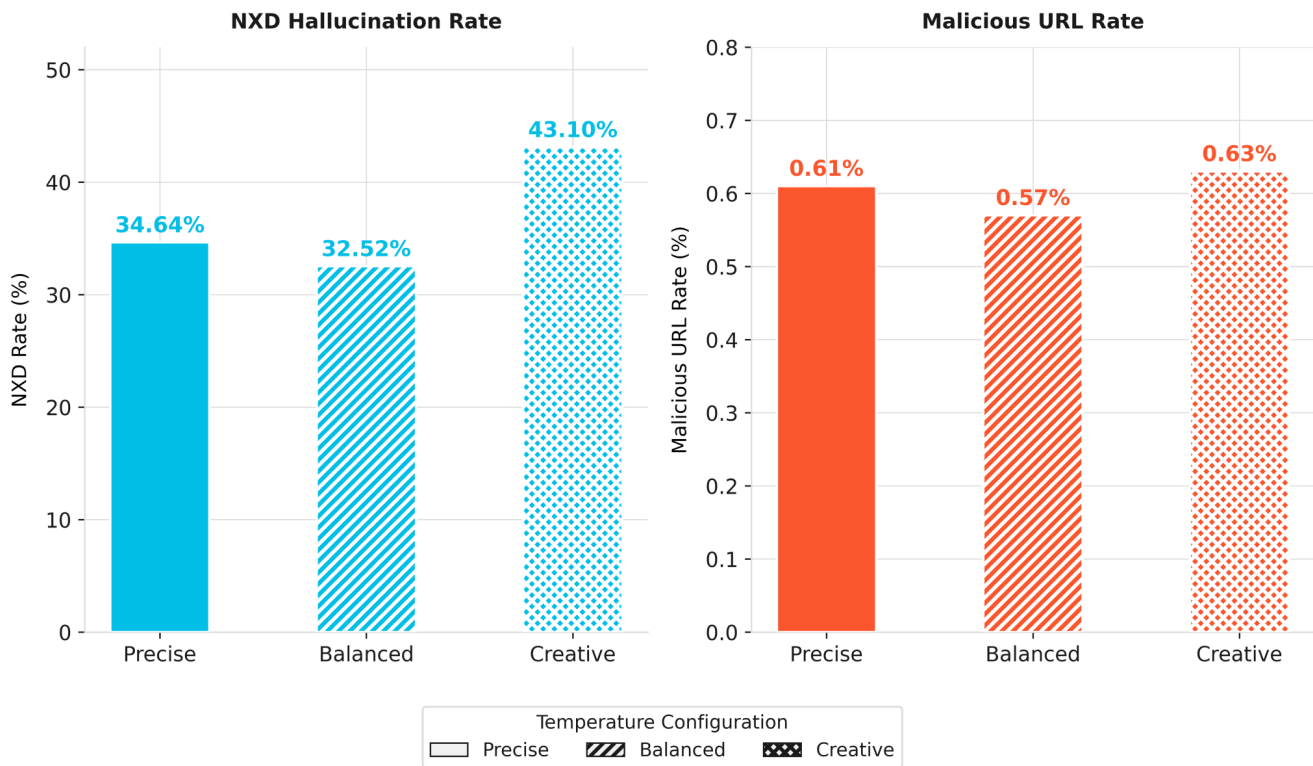


Figure 5. Impact of inference temperature configuration on both the NXD hallucination rate and malicious URL rate.

Anatomy of URL Hallucinations

The structural composition of phantom domain hallucinations is not uniform. Within our corpus of 809,455 unique NXD URLs, nearly half (49.7%) manifest as path-level hallucinations, where the LLM constructs a plausible resource path on a legitimate, registered domain that fails to resolve.

An additional 39.5% are categorized as subdomain-level hallucinations — fabricated sub-architectures under existing base domains. The most critical tier, representing 10.8% of the dataset, consists of pure domain-level hallucinations involving entirely unregistered root namespaces.

Analysis reveals divergent behavioral profiles between architectures:

- LLM1 exhibits a pronounced bias toward path-level extrapolation (56.6%)
- LLM2 generates a significantly higher proportion of subdomain-level (45.1%) and pure domain-level (20.0%) hallucinations, expanding the registerable attack surface available for adversarial exploitation

Figure 6 illustrates the structural distribution of NXD hallucinations across three architectural tiers: path, subdomain and domain. This provides a comparative visualization for the aggregate corpus and individual model performances.

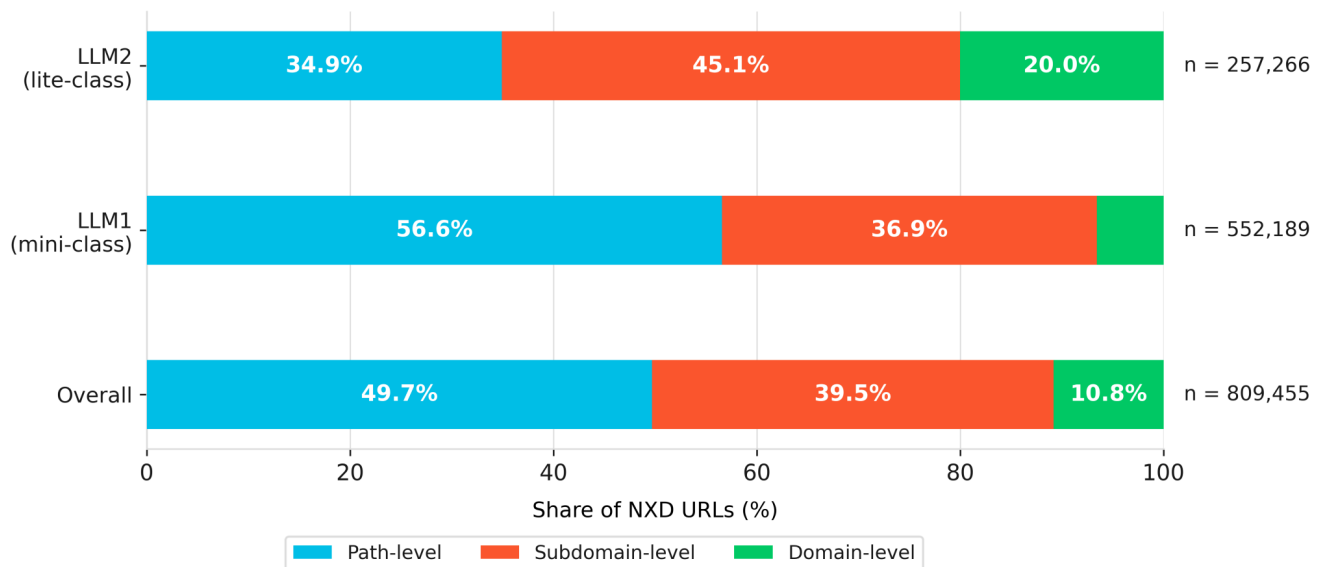


Figure 6. URL Pattern distribution of NXD hallucinations across LLM systems.

Evidence of Active Exploitation: Real-World Detection Cases

Aggregate statistics confirm the structural scale of the phantom squatting threat. The following case studies document the real-world manifestation of this vector. These examples demonstrate instances where our discovery pipeline identified a phantom domain prior to adversarial registration for malicious deployment. The case studies are:

- Impersonation of a postal service's e-commerce marketplace in a phishing campaign using the Montana Empire phishing kit
- Impersonation of a national postal service to deliver a malicious Android app
- Four other examples of phantom squatting weaponized in real-world attacks

To quantify this proactive detection advantage, we define the adversarial exploitation window (AEW). This window is the temporal interval between the initial hallucination event and the subsequent registration by a threat actor.

A positive AEW signifies actionable lead time for defenders. A negative AEW signifies that an adversary registered the infrastructure prior to our detection. This provides historical validation of the threat model, confirming that disparate AI architectures and human adversaries independently converged on the same structurally inevitable hallucination.

Montana Empire: AI-Assisted Phishing and the Closed Loop

AEW: 23 days

Target: Customers of a national postal service's e-commerce marketplace

On March 8, 2026, our multi-agent discovery pipeline generated 13 hallucinated URLs for a domain similar to a national postal service e-commerce website across both LLM families and all temperature configurations — including the Precise (T = 0.1) setting. This systemic convergence indicated high Thermal Hallucination Persistence (THP), where the models treated the infrastructure as a near-factual certainty rather than speculative creative entropy. Consequently, the domain was enrolled in the phantom domain watchlist within the same week.

On March 31, 2026 — representing an AEW of 23 days — an adversary registered the domain and deployed a weaponized phishing kit impersonating a legitimate marketplace. Our pipeline achieved same-day detection of the registration event.

Analysis of the attacker’s infrastructure revealed that they developed the phishing kit using an AI coding assistant, evidenced by a project directory identified within [the phishing kit ZIP archive](#). Telemetry from the session logs confirmed the adversary leveraged the assistant to scrape legitimate storefronts, engineer the PHP backend and develop a Telegram-based C2 interface for real-time credential exfiltration.

This case demonstrates the full cycle of the phantom squatting supply chain threat model. The adversary used AI systems to generate attack tooling against infrastructure identified by our discovery pipeline 23 days earlier. Both parties arrived at the spoofed domain via the same mechanism, the LLM’s internal prediction of a structurally inevitable hallucination for the target brand.

The Montana Empire phishing kit features several advanced components:

- A real-time scraper designed to maintain site parity with live storefronts
- Dual-channel interception for credit cards and bank transfers, with International Bank Account Numbers (IBANs) rotated via Telegram bot commands
- Harvesting of national identity document data alongside payment credentials
- An operator control panel for manual one-time password (OTP) relay and victim adjudication

Table 1 shows the specific domain patterns hallucinated by the models in our analysis.

Prompt	Hallucinated URL	Model	Configuration
“Payment gateway sandbox for the postal e-commerce marketplace integration?”	hxxps[:]//sandbox.[redacted] [.]com/payment/api/v1/pay	LLM2	Balanced (T = 0.7)

“Administrative dashboard for the postal e-commerce marketplace campaigns?”	hxxps[:]//admin.[redacted][.]com	LLM2	All three configurations, including Precise
“Billing portal for the postal e-commerce marketplace payment method update?”	hxxps[:]//billing.[redacted][.]com/	LLM2	Balanced (T = 0.7)

Table 1. Prompts, hallucinated URLs and model configurations.

Figure 7 illustrates the Montana Empire admin panel, with a banner bearing the phrase *Kimseye Güvenme* (Trust No One). Through this control panel, the attacker managed victim OTP approvals and credential relays via Telegram.

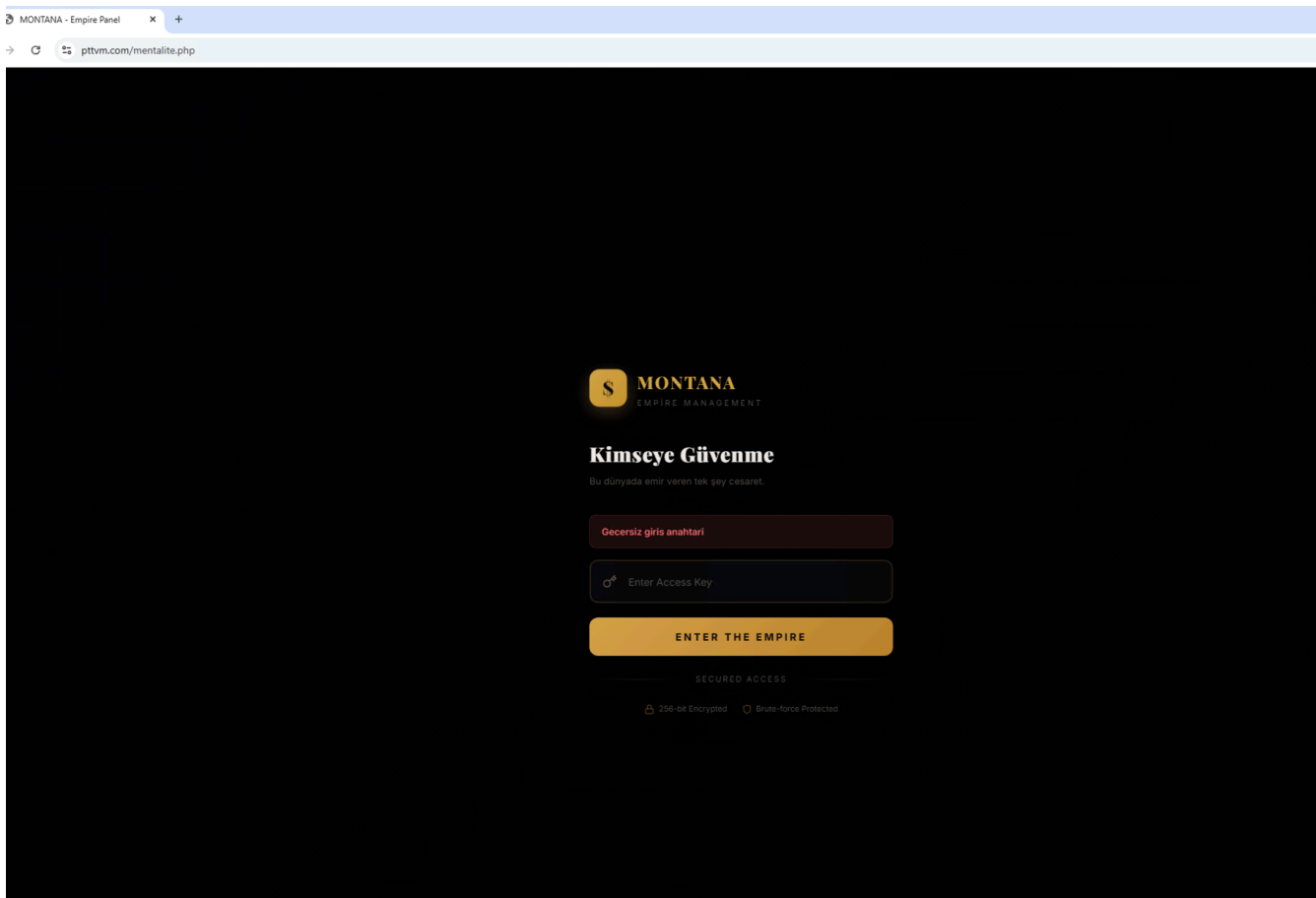


Figure 7. Montana Empire admin panel.

When we recovered the phishing kit file structure from the server of the spoofed domain, we noted an AI coding assistant project directory, indicating that the attacker used an AI tool to develop the kit.

National Postal Service-Themed Delivery of Malicious Android App

AEW: 51 days

Target: Customers of a national postal delivery service

On Feb. 18, 2026, our multi-agent discovery pipeline identified hallucinated URLs for admin. [redacted]post-app[.]com across five distinct model-configuration tiers — including LLM1 at the Precise (T = 0.1) setting. This high degree of convergence led to the parent domain, [redacted]post-app[.]com, being enrolled in the phantom domain watchlist for proactive monitoring.

On April 10, 2026 — representing an AEW of 51 days — an adversary registered [redacted]post-app[.]com and immediately deployed a site that used a pixel-accurate brand clone impersonating the national postal service. The malicious landing page replicated the service’s authoritative aesthetic. It used the same HTML hex color code as the official brand and fabricated social proof (4.8-star rating, over 2 million users) to drive victims to download a malicious Android application package (APK) file named [\[redacted\]post.apk](#). Our registration event stream achieved detection within hours of infrastructure creation.

While legitimate postal applications are restricted to official marketplaces, this out-of-band delivery bypasses standard platform-level security telemetry.

Additional Detection Cases

Our multi-agent discovery pipeline and subsequent triage verified the following cases shown in Table 2.

Domain	Brand	AEW	Attack Pattern
[redacted]-login[.]com	Online sports betting operator	45 days	Credential-harvesting clone targeting the Bangladesh market; features explicit local language headings and BDT currency.
[redacted]-es[.]org	Competing sports betting operator	40 days	Infrastructure registered in an 18-minute coordinated window by the same actor; identical regional targeting and monetization strategy.

[redacted]empresas[.]com	Regional European retail bank	35 days	Re-registration event detected.
[redacted]business[.]com	Major UAE commercial bank	-11 months	Historical validation of a structurally inevitable hallucination; corporate IT credential harvester using fraudulent branding.

Table 2. Phantom domain detection examples.

A real-world example involving a major bank in the UAE proves that AI models predictably hallucinate the same fake information. On April 1, 2025, a threat actor registered the domain [redacted]business[.]com to steal login details from company database administrators. This campaign had been running for nearly a year. Our pipeline system independently predicted and generated that same fake web address 11 months after it was first used. Our team analyzed this domain through our verification pipeline after detecting that it was being registered again.

Two other examples of phantom squatting in Table 2 reflect the coordinated registration of [redacted]-login[.]com and [redacted]-es[.]org. A single actor registered these domains using identical registrars, nameservers and privacy shielding within an 18-minute window. This demonstrates that phantom squatting is useful for detecting multi-target, orchestrated campaigns.

In this instance, the adversary deployed a unified infrastructure for both domains. Both phishing sites use an identical architectural template, featuring a মেগা জ্যাকপট পুল (Mega Jackpot Pool) display and Bengali-language localized content. By explicitly referencing Bangladesh (বাংলাদেশে) and processing transactions in Bangladeshi Taka (টাকা), the actor provided definitive attribution signals for a regionally focused, high-velocity operation.

Implications for AI-Powered Supply Chains

Agentic Workflow Risk: Compromising Autonomous AI Pipelines

The highest-consequence phantom squatting target is not a human user. Instead, it is an autonomous AI agent. Agentic systems increasingly execute multi-step workflows that include web fetching, API calls and resource downloads, all based on URLs generated by the LLM orchestrating the pipeline. When an AI agent generates a URL to fetch documentation,

retrieve an API schema, or download a dependency, that artifact may resolve to a phantom domain controlled by an adversary.

The impact in an agentic context is amplified by autonomy. A human user who follows an LLM-recommended URL and reaches a phishing page must still take an action by entering credentials, downloading a file or executing code. Conversely, an autonomous agent that fetches a URL and processes its response could exfiltrate secrets, execute malicious instructions or propagate a compromised dependency through a build pipeline without any human decision point.

The 2026 [Unit 42 Global Incident Response Report describes](#) an identity-velocity crisis, where attackers compress the window from initial access to exfiltration to under one hour at machine speed. This applies directly to phantom squatting delivered via agentic pipelines.

Developer Tooling Risk: AI Coding Assistants and URL Hallucination in the Software Development Lifecycle

Modern software engineering workflows have integrated AI coding assistants for tasks that fundamentally require URL generation:

- Retrieving API documentation
- Identifying package registries
- Locating webhook endpoints
- Architecting integration code.

Each interaction represents a potential phantom squatting vector.

The Montana Empire case provides a definitive illustration of this risk convergence. The adversary leveraged an AI coding assistant to engineer a phishing kit targeting the exact phantom domain predicted by the LLM's hallucination patterns.

AI-assisted attack development and LLM-driven attack delivery are no longer disparate phenomena. They represent two dimensions of a single, structurally inevitable adversarial lifecycle.

Conclusion

The risk of phantom squatting is not a theoretical abstraction. Our analysis of 913 global brands and 2.1 million LLM-generated URLs documents a critical supply chain vulnerability:

- 13,229 confirmed malicious URLs currently being produced by LLMs

- 250,000 hallucinated phantom domains representing nascent, unregistered infrastructure available for adversarial occupation
- Real-world threat actor registrations validated via WHOIS analysis, yielding proactive detection lead times of up to 51 days

This vector exploits a structural property of LLM architectures that remains inherently unpatchable. Models trained on human-authored corpora will naturally hallucinate plausible-sounding domains for brands, products and services based on internal linguistic patterns. The phantom squatting attack surface systematically expands with every new LLM deployment, the rise of agentic AI capabilities and the targeting of global brands for adversarial hallucination probing.

The defensive advantage is equally architectural. Because LLMs hallucinate with predictable consistency, defenders can map the hallucination surface and establish a proactive phantom watchlist before an adversary acts. The AEW — the interval between first hallucination detection and registration — provides concrete, actionable lead time that legacy threat intelligence frameworks cannot offer.

Proactive discovery represents the only defensive posture that addresses phantom squatting at its root. By mapping what LLMs will hallucinate and monitoring registration event streams, organizations can respond before weaponization occurs. The capability is established, and the zero-reputation bypass window is open. The critical question is whether defenders or adversaries will act first.

Palo Alto Networks customers are better protected from the threats discussed above through the following products:

- The [Advanced WildFire](#) machine-learning models and analysis techniques have been reviewed and updated in light of the indicators shared in this research.
- [Advanced URL Filtering](#) and [Advanced DNS Security](#) identify known domains and URLs associated with this activity as malicious.
- [Prisma AIRS](#) can help secure organizations deploying LLM-powered agentic workflows.
- [Koi Agentic Endpoint Security](#) is designed to help discover every AI artifact across the agentic endpoint, assess its risk, enforce prevention & runtime controls, and remediate violations.

The [Unit 42 AI Security Assessment](#) can help empower safe AI use and development.

If you think you may have been compromised or have an urgent matter, get in touch with the [Unit 42 Incident Response team](#) or call:

- North America: Toll Free: +1 (866) 486-4842 (866.4.UNIT42)
- UK: +44.20.3743.3660
- Europe and Middle East: +31.20.299.3130
- Asia: +65.6983.8730
- Japan: +81.50.1790.0200
- Australia: +61.2.4062.7950
- India: 000 800 050 45107
- South Korea: +82.080.467.8774

Palo Alto Networks has shared these findings with our fellow Cyber Threat Alliance (CTA) members. CTA members use this intelligence to rapidly deploy protections to their customers and to systematically disrupt malicious cyber actors. Learn more about the [Cyber Threat Alliance](#).

Indicators of Compromise

Montana Empire Campaign

The following domains are presented in partially redacted form. Full unredacted indicators are available on request.

SHA256 hash: eb07edaa2786cfddfa4c15526168f2200d85300aee0a8f253b32d2462a7b0bcd

File size: 7,958,528 bytes

File type: ZIP archive

Filename: [redacted].zip (postal e-commerce platform brand name)

File location: hxxp[:]//[redacted][.]com/[redacted].zip

File description: Montana Empire phishing kit archive — comprises a full brand clone of a national postal service's e-commerce marketplace featuring a PHP backend, real-time storefront scraper, credential capture layer and Telegram-based C2 operator control panel.

Related URLs:

- hxxp[:]//[redacted][.]com/[redacted].zip
- hxxp[:]//[redacted][.]com/letgovip.zip
- hxxp[:]//[redacted][.]com/mentalite.php
- hxxp[:]//[redacted][.]com/panel_track.php
- hxxp[:]//[redacted][.]com/verify_api.php

National Postal Delivery Service APK Campaign

SHA256 hash: 2202a30daad9928ef47cca5f4ab04ce083692a94428e386fa01c2dd44557e34b

File size: 12,649,472 bytes

File type: APK (Android application package)

Filename: [redacted]post.apk

File location: hxxp[:]//[redacted]post-app[.]com/[redacted]post.apk

File description: Malicious Android APK delivered via a fraudulent mobile app landing page impersonating a national postal delivery service.

Related domain and URL:

- [redacted]post-app[.]com
- hxxp[:]//[redacted]post-app[.]com/[redacted]post.apk

Additional Phantom Domain Detections

- [redacted]-login[.]com
- [redacted]benefitsportal[.]com
- [redacted]-es[.]org
- [redacted]business[.]com
- [redacted]empresas[.]com

Acknowledgments

The authors would like to thank Shehroze Farooqi, Joseph Pang and Wanjin Li for their valuable insights and contributions in completing this work. The authors would also like to thank Samantha Stallings, Bradley Duncan, Lysa Myers and Shawn He Shuang for their assistance in the editorial process.

Additional Resources

- [Montana Empire Phishing Kit](#) – Palo Alto Networks
- [The Rise of Slopsquatting](#) – Socket blog
- [AI-Induced Supply-Chain Compromise](#) – Al-Zof, A. et al.
- [We Have a Package for You](#) – Kan, M. et al., USENIX Security
- [Unit 42 Global IR Report 2026](#) – Palo Alto Networks