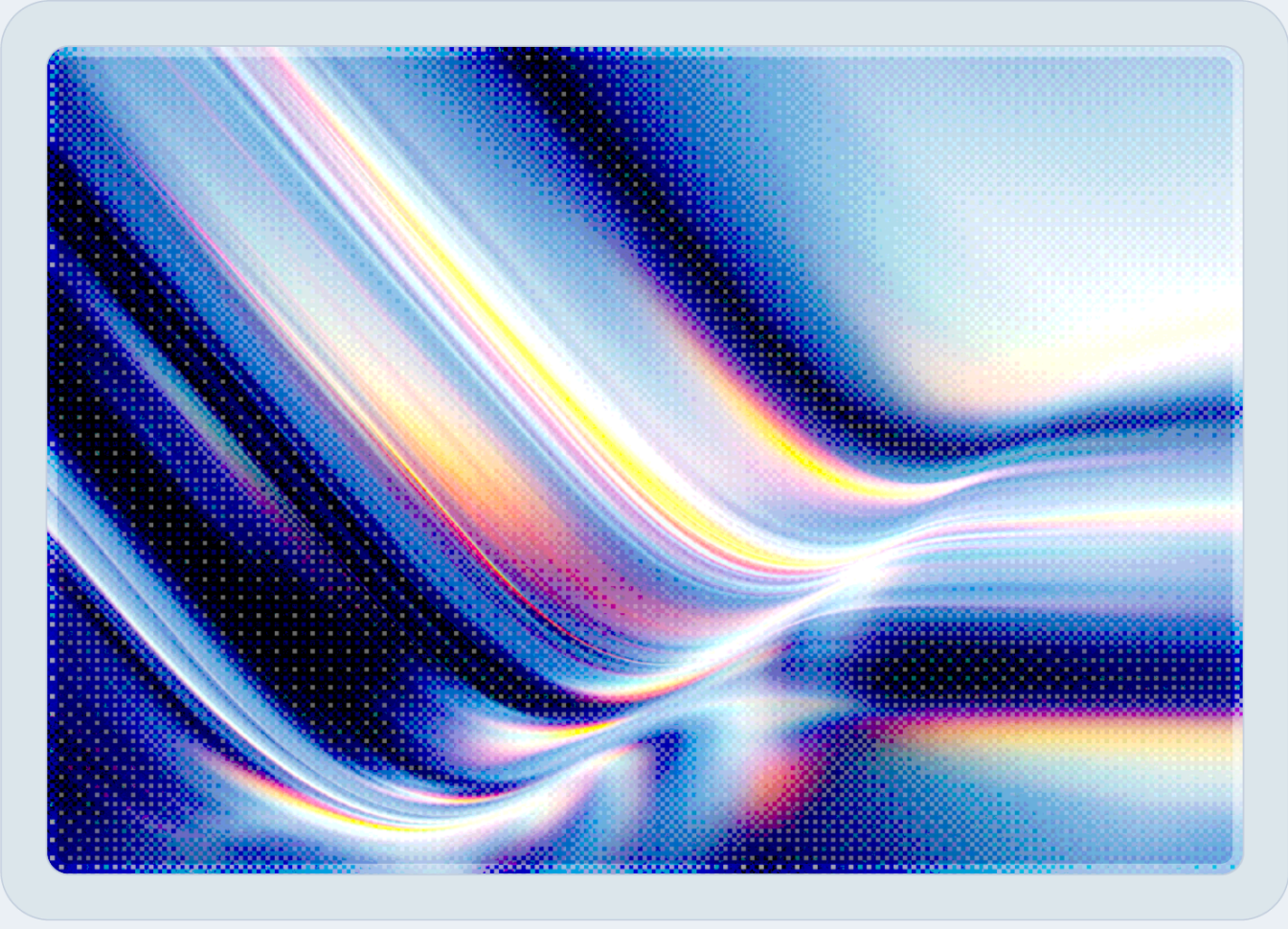
CSAI Foundation | Cloud Security Alliance

# Phantom Squatting: Attackers Exploit AI-Hallucinated Domains

How LLM Hallucination Consistency Creates Predictable Phishing Infrastructure

2026-07-01

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

## Key Takeaways

- Large language models (LLMs) consistently generate plausible but nonexistent web domains when responding to user queries about well-known brands. This hallucination is not random — models produce the same fabricated domains repeatedly, giving attackers a predictable list of registration targets.
- Palo Alto Networks Unit 42 analyzed 913 global brands and generated 2.1 million URLs from 685,339 adversarial prompts, finding 809,455 non-existent domain URLs (37.28%) and approximately 250,000 unregistered phantom domains available for immediate adversarial registration [1].
- Two confirmed real-world campaigns — including the Montana Empire phishing kit — demonstrate that attackers are already registering AI-hallucinated domains and standing up credential-harvesting infrastructure, with Unit 42 documenting advance warning windows of 23 to 51 days between hallucination detection and attacker registration [1].
- Because hallucination patterns are consistent and measurable, defenders can proactively identify which domains a model is likely to fabricate and monitor domain registration streams for early warning — a meaningful shift from purely reactive threat intelligence.
- This threat is closely related to slopsquatting, which targets AI-hallucinated software package names rather than web domains. Both attack classes exploit the same root cause: LLMs generating authoritative-sounding recommendations for resources that do not exist.
- Organizations deploying AI assistants, agentic workflows, or AI-powered customer tools should treat model-generated URLs as unverified and implement controls that prevent autonomous navigation to model-suggested domains without human review.

---

## Background

The hallucination problem in large language models is well documented [7]: models regularly generate confident, plausible-sounding content that is factually incorrect. When that content takes the form of web addresses — URLs that a model presents as belonging to a trusted brand — the consequences

extend well beyond a minor factual error. A user who asks an AI assistant for the web portal of a bank, postal service, or government agency and receives a hallucinated URL does not know the address is fabricated. If an attacker has already registered that domain and populated it with a convincing clone of the real site, the user proceeds through a phishing funnel that the AI itself initiated.

This attack technique, named phantom squatting by Palo Alto Networks Unit 42, was documented in a July 2026 research publication following an extended period of systematic measurement [1, 2]. Phantom squatting belongs to the broader category of domain squatting attacks — registering a domain name similar to a legitimate brand's to intercept misdirected traffic — but it introduces a qualitatively different exploitation path: instead of depending on user error to misdirect traffic, it exploits the authoritative outputs of a system users explicitly trust for navigation guidance. Traditional domain squatting and typosquatting depend on user keystroke errors or memorization mistakes. Phantom squatting, by contrast, is machine-generated: the AI model itself supplies the bad address, with the authority and confidence that users have come to associate with AI-provided information.

The threat shares a conceptual ancestor with slopsquatting, a supply chain attack that emerged from the same hallucination dynamic applied to software packages. In slopsquatting, LLMs generating code produce import statements referencing package names that do not exist in public registries; attackers register those names on npm, PyPI, or other repositories and embed malware in the packages developers then install [3, 7, 8]. CSA's April 2026 research note on slopsquatting documented that 19.7% of AI-generated code samples contained fabricated package names, and that 61% of hallucinated names reappeared consistently across prompt runs, making them reliable exploitation targets [4]. Phantom squatting extends this attack surface from the software supply chain to the open web, targeting the end users and automated agents that consume AI-generated navigation guidance rather than the developers who consume AI-generated code.

---

# Security Analysis

## Measurement and Scale

Unit 42 conducted a structured measurement study spanning 913 brands across technology, finance, healthcare, e-commerce, government, gambling, and logistics sectors [1]. Researchers executed 685,339 adversarial prompts against two LLM architectures — an enterprise mini-class model and a frontier lite-class model — under three temperature settings (precise at T=0.1, balanced at T=0.7, and creative at T=1.5). The resulting corpus of 2.1 million URLs was evaluated for registration status, malicious activity, and cross-model hallucination agreement.

The scale of hallucination was substantial. Of the 2.1 million generated URLs, 809,455 (37.28%) pointed to non-existent domains or paths. Within that population, Unit 42 identified approximately 250,000 unregistered phantom domains that adversaries could register immediately, 13,229 URLs already confirmed as hosting malicious content (0.61%), and 41,313 URLs rated as high-risk (1.90%) [1]. The structural breakdown of hallucinated URLs showed that 49.7% were path-level fabrications on legitimately registered root domains, 39.5% were subdomain-level fabrications on unregistered sub-architectures, and 10.8% were pure domain-level fabrications where the entire root domain was invented [1]. This last category — entirely fabricated domain registrations — is the most directly actionable for attackers, as the root domain itself can be claimed.

Temperature configuration had a measurable but secondary effect on hallucination frequency. Creative-mode prompting produced non-existent domain URLs at a rate of 43.10%, compared to 34.64% for precise-mode prompting — a meaningful difference, but not one that eliminates the risk at lower temperature settings [1]. More significantly, malicious URL rates remained consistent across model architectures (0.56%–0.64%), indicating that the relationship between hallucinated domains and adversarial content is a structural property of model output rather than a consequence of entropy-driven variation. Lowering temperature settings reduces hallucination volume but does not eliminate adversarial exposure.

The two LLM architectures diverged in hallucination profile. The mini-class model produced non-existent domain URLs at a rate of 44.6%, compared to 27.5% for the lite-class model. These rates also differed in hallucination structure: the mini-class model biased toward path-level fabrications (56.6%), while the lite-class model generated proportionally more subdomain-level (45.1%) and domain-level (20.0%) hallucinations [1]. The practical implication is that organizations deploying smaller, faster models as AI assistants may face a higher absolute volume of hallucinated web addresses than those using larger frontier models.

## The Phantom Squatting Attack Lifecycle

Unit 42 describes the phantom squatting attack in four phases [1]. In the discovery phase, adversaries systematically query LLMs with prompts designed to elicit brand-related URLs. Because models hallucinate consistently — generating the same fabricated domain repeatedly across independent queries — attackers can map the hallucination landscape for a target brand with relative confidence. In the acquisition phase, threat actors register unregistered phantom domains; sophisticated actors would logically prioritize those that appear across multiple models and prompt configurations, since cross-model hallucination consensus signals both high traffic potential and broad user exposure — though this specific prioritization behavior has not yet been documented in confirmed attacker tooling. In the lure phase, users or autonomous AI agents that issue a triggering query receive an authoritative

recommendation directing them to the attacker-controlled domain. No phishing email, malicious advertisement, or social engineering is required; the AI model's own output redirects the user to attacker-controlled infrastructure. In the bypass phase, the newly registered domain exploits zero-reputation status: it has no blocklist entries, no malicious history, and no threat intelligence associations, because it was clean at the moment of registration.

This zero-reputation bypass is the attack's structural advantage over conventional phishing. Traditional phishing domains accumulate detection signals over time as users report them, security vendors analyze them, and threat feeds propagate the results. A phantom domain registered specifically to receive AI-directed traffic generates no prior signal whatsoever. By the time behavioral threat intelligence synchronizes on the domain's malicious activity, a population of users has already been delivered there by a system they consider trustworthy.

## Confirmed Real-World Campaigns

The research documented two confirmed phantom squatting campaigns against national postal service brands, representing two of the clearest examples of the attack lifecycle in practice [1].

In the first case, Unit 42's pipeline identified 13 hallucinated URLs for a domain resembling a national postal service's e-commerce marketplace on March 8, 2026. Twenty-three days later, on March 31, an attacker registered the exact domain and deployed a phishing kit named Montana Empire. The kit included a full brand clone, a PHP backend, a real-time storefront scraper, credential capture logic, and a Telegram-based command-and-control channel with support for one-time password relay and victim handling. Post-incident forensic review of leftover project files and session logs confirmed that the attacker had developed the kit using an AI coding assistant — a closed-loop scenario in which AI systems contributed to both the delivery mechanism (hallucinating the domain) and the construction of the attack infrastructure [1, 5].

In the second case, researchers flagged a hallucinated postal service administration subdomain on February 18, 2026, when five distinct model-configuration combinations independently generated it. Fifty-one days later, on April 10, an attacker registered the parent domain and distributed a malicious Android APK using a pixel-perfect brand clone that replicated ratings, user reviews, and UI elements from the legitimate application [1]. Additional cases documented in the study include credential harvesters targeting sports betting operators (40–45 day advance windows) and a European retail bank credential-harvesting clone (35 days). One UAE commercial bank case validated the hallucination prediction methodology against historical registration data extending eleven months [1].

## Malicious Content Distribution

Among the 13,229 confirmed malicious URLs identified in the study corpus, malware distribution dominated: 67.2% of malicious URLs led to drive-by downloads or exploit kits, 16.2% hosted credential-harvesting phishing pages, 13.7% delivered grayware including adware and potentially unwanted programs, and 3.0% served as command-and-control infrastructure [1]. This distribution suggests that phantom squatting as currently observed is not exclusively a phishing technique — adversaries are exploiting AI-directed traffic to distribute malware broadly, and the phishing component is present but secondary to malware delivery by volume.

## Agentic AI Amplification

The threat takes on additional dimensions in agentic AI deployments, where models are connected to the web and authorized to navigate autonomously on behalf of users. An agentic assistant directed to research competitors, retrieve pricing information, verify account status, or perform e-commerce transactions may follow the web addresses it generates without the human observation that might catch a suspicious URL — particularly if runtime guardrails do not constrain outbound navigation. Where a human user might notice a suspicious domain in a browser's address bar, an agentic pipeline navigating programmatically has no equivalent checkpoint. Any credentials, payment instruments, or sensitive data exchanged during an agentic session on a phantom domain are captured by the attacker without the victim ever seeing the URL.

This amplification risk is concrete, not speculative — the same hallucination mechanics apply identically to agentic pipelines, and the exposure grows as agents gain authorization to handle sensitive operations. AI agents operating in enterprise environments are increasingly authorized to access authenticated systems, initiate financial transactions, and handle personal data. The Montana Empire campaign targeted payment credentials and national identity documents. An agentic financial assistant authorized to retrieve account statements from a bank's portal — and directed to a phantom domain instead — faces a structurally identical credential exposure pathway.

---

# Recommendations

## Immediate Actions

Security teams should begin treating AI-generated URLs as untrusted by default. No organization should allow AI assistants or agentic systems to navigate to model-generated web addresses without URL verification against authoritative sources. For brand-facing applications that use AI to provide web guidance to users, the responsible disclosure window may be narrow: Unit 42's data suggests that high-confidence hallucination targets can be registered within weeks of discovery. Organizations with significant brand presence in finance, logistics, healthcare, or e-commerce should initiate a hallucination audit of the AI systems they deploy, querying those systems with prompts likely to elicit brand-related URLs and evaluating whether any of the resulting addresses are unregistered and registrable by adversaries.

For security operations teams, monitoring of domain registration streams for brand-adjacent strings should incorporate hallucination-pattern analysis. The brands most commonly hallucinated by the LLMs in the Unit 42 study spanned multiple verticals, and the predictability of hallucination outputs means that a targeted watchlist of phantom domain candidates can be constructed and fed into existing domain monitoring pipelines. The 23–51 day advance warning windows documented in the real-world cases suggest that proactive hallucination mapping can yield actionable lead time.

## Short-Term Mitigations

Organizations deploying AI assistants in customer-facing or employee-facing contexts should implement URL verification as a mandatory step before any model-suggested link is presented to users or navigated to by automated agents. Verification should check registered domain status, registration date (recently registered domains — within 90 days as a starting point, calibrated to local threat context — warrant elevated scrutiny), and consistency with the brand's confirmed domain portfolio. Where AI tools surface external URLs in responses, those links should be intercepted and validated against threat intelligence before being made clickable.

For agentic AI systems authorized to access the web, runtime guardrails should enforce domain allowlists for sensitive operations involving authentication, payment, or personal data. An agentic system navigating to an authentication portal for a financial institution should only be permitted to proceed to confirmed, pre-approved domains from that institution's registered namespace. URL pattern matching at the allowlist level, rather than relying on real-time threat intelligence alone, closes the zero-reputation bypass window that phantom squatting exploits.

Development teams using AI coding assistants should apply the slopsquatting mitigation discipline [4] as an analogous control: just as AI-generated import statements should be verified against package registries before installation, AI-generated web URLs should be verified against authoritative brand domain registries before use in application logic or documentation.

### Strategic Considerations

At the organizational level, brand protection programs should integrate LLM hallucination monitoring as a dedicated threat category alongside traditional domain monitoring. The tooling required is tractable: the same models that generate hallucinated domains can be systematically queried to surface the most persistent and cross-model-consistent phantom domain candidates, which become the highest-priority registration monitoring targets. Security vendors have begun offering phantom domain watchlist capabilities, and organizations with large brand footprints should evaluate these offerings against the demonstrated risk.

Enterprises procuring AI-powered tools and autonomous agents from vendors should incorporate hallucination safety and URL handling into security assessments and procurement requirements. Vendor AI security assessments should ask specifically how the product handles model-generated web addresses, whether outbound navigation by AI agents is constrained to verified domains, and whether the vendor monitors its models for hallucination-driven external link generation. The STAR for AI program provides a framework for this kind of vendor evaluation.

The predictability of LLM hallucination patterns creates a long-term opportunity for defensive intelligence that does not exist in most conventional threat categories. Because models hallucinate the same fabricated domains consistently, and because cross-model consensus on a hallucinated domain is a strong predictor of adversarial registration interest, a coordinated hallucination intelligence capability — shared across the security community — could shift the defense-attacker balance on this threat class. CSA's existing community threat sharing infrastructure could provide a mechanism for distributing hallucination-derived phantom domain intelligence, complementing existing domain reputation feeds with predictive pre-registration signals.

---

## CSA Resource Alignment

The phantom squatting threat engages several dimensions of CSA's AI security framework portfolio.

**MAESTRO** (Multi-Agent Environment, Security, Threat, Risk, and Outcome), CSA's agentic AI threat modeling framework, is directly applicable to the phantom squatting risk in agentic deployments [6]. MAESTRO Layer 2 — AI Reasoning and Logic — addresses threats arising from model outputs that are unreliable or adversarially conditioned. The consistent production of hallucinated web addresses is a reasoning-layer vulnerability: the model generates plausible outputs that are factually wrong, and downstream agent behaviors that treat model output as authoritative transform that reasoning failure into an active attack surface. Threat modeling exercises using MAESTRO for AI agents authorized to navigate the web should explicitly include phantom domain generation as a threat scenario.

**AI Controls Matrix (AICM)** addresses AI supply chain security and runtime integrity. The Supply Chain Management domain within AICM provides control objectives relevant to verifying that AI-generated content pointing to external resources is validated before use. Organizations using the AICM for internal AI governance should map phantom squatting risk to controls governing model output verification and agentic action authorization, particularly for controls in the Application Security and Identity and Access Management domains where unverified model-generated URLs could enable authentication attacks.

**STAR for AI** provides the vendor assessment mechanism that organizations should apply when procuring AI tools with web-navigation or URL-generation capabilities. The phantom squatting risk should be incorporated into STAR-based AI procurement reviews as a specific inquiry category, requesting evidence of vendor controls over hallucinated URL generation and agentic navigation guardrails.

**CSA Research Note: Slopsquatting** (April 2026) [4] established the foundational analysis of LLM hallucination consistency as a supply chain attack vector. Phantom squatting represents a web-domain extension of the same threat model, and the two notes should be read together to understand the full scope of hallucination-driven adversarial exploitation. Controls recommended for slopsquatting — including treating AI-generated dependencies as untrusted, enforcing verification before use, and monitoring hallucination patterns — apply directly, by analogy, to phantom squatting in the web domain space.

# References


[1] Palo Alto Networks Unit 42. "Phantom Squatting: AI-Hallucinated Domains as a Software Supply Chain Vector." Palo Alto Networks Unit 42, July 1, 2026.

[2] The Hacker News. "Phantom Squatting Uses AI-Hallucinated Domains for Phishing and Malware." The Hacker News, July 1, 2026.

[3] Sonatype. "PhantomRaven: npm Malware Evolves Again." Sonatype, October 2025.

[4] Cloud Security Alliance AI Safety Initiative. "Slopsquatting: AI Package Hallucinations and Software Supply Chain Risk." CSA Labs, April 19, 2026.

[5] Cyberpress. "Montana Empire Phishing Kit Abuses AI-Hallucinated Domain to Steal Credentials." Cyberpress, 2026.

[6] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[7] CSO Online. "AI Hallucinations Lead to a New Cyber Threat: Slopsquatting." CSO Online, April 14, 2025.

[8] Socket Security. "Slopsquatting: How AI Hallucinations Are Fueling a New Class of Supply Chain Attacks." Socket.dev, April 8, 2025.