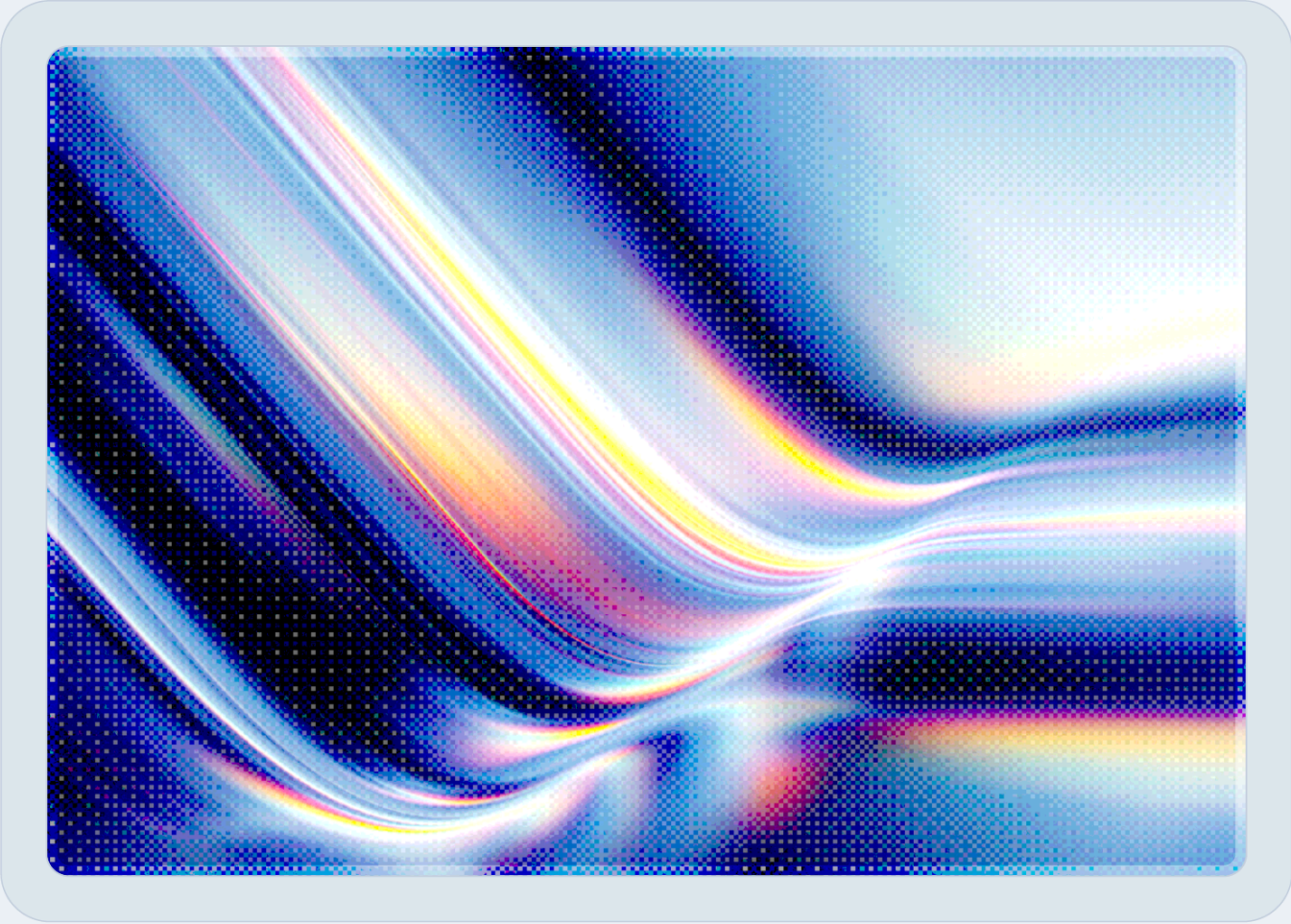
CSAI Foundation | Cloud Security Alliance

Phantom Squatting: AI Hallucinated Domains as Phishing Infrastructure

2026-07-02

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Large language models frequently hallucinate domain names for real organizations, generating URLs that look authoritative but point to unregistered or nonexistent infrastructure. Adversaries deliberately probe these models, register the resulting phantom domains before defenders notice, and weaponize them for phishing and malware delivery.
- Research published by Palo Alto Networks Unit 42 in June 2026 analyzed 913 global brands across 685,339 adversarial prompts and two LLM families, generating 2.1 million URLs. Of those, 809,455 pointed to non-existent domains — and 13,229 had already been independently flagged as malicious by threat intelligence feeds [1].
- Approximately 250,000 hallucinated domains remain unregistered and available for immediate adversarial registration, representing a large, measurable, and preventable attack surface [1].
- Phantom squatting is a structural threat amplified by agentic AI deployments. Autonomous agents that act on LLM-generated URLs without verification create attack paths requiring no human decision point between model output and compromise.
- Proactive hallucination mapping provides a measurable defensive advantage. Unit 42's monitoring pipeline demonstrated lead times of 18 to 51 days between hallucination detection and adversarial domain registration, giving defenders a predictive window that reactive blocklists cannot offer [1].

Background

From Typosquatting to Slopsquatting to Phantom Squatting

Domain-based impersonation attacks have evolved alongside the internet itself. Typosquatting exploits human transcription errors by registering minor misspellings of well-known domains. Combosquatting appends legitimate-seeming words to brand names. Both techniques depend on predicting human behavior — the attacker bets that a user will mistype or misread a URL.

A newer variant, slopsquatting, emerged from AI code generation workflows. When developers use AI assistants to scaffold code, those models sometimes recommend software packages that do not exist in any legitimate registry. Security researchers documented cases where up to 20% of AI-generated code samples referenced non-existent package names, with 43% of hallucinated names recurring consistently across multiple model invocations — making them predictable enough to preemptively register [2][6]. CSA's AI Safety Initiative documented this supply chain vector in detail in April 2026 [3].

Phantom squatting extends the same logic from software packages to web domains. Rather than targeting package registries like npm or PyPI, it targets the broader infrastructure of URLs that LLMs generate when answering questions about organizations, services, and products. The underlying mechanism is structurally analogous: an AI system produces a plausible-sounding artifact it treats as factual, a human or autonomous agent acts on it, and an adversary positioned at that artifact captures the traffic.

Why LLMs Hallucinate URLs

Language models are trained to generate statistically probable outputs, not verified facts. When a model is asked for the official website of a company's customer portal, its developer API endpoint, or the download location of a specific software tool, it assembles a URL from patterns in its training data rather than consulting a live registry. The result frequently looks indistinguishable from a legitimate address: correct brand name, plausible subdomain structure, appropriate top-level domain. The model may express this output with the same confidence it would use for a verifiable fact.

This behavior is not limited to obscure or niche queries. Unit 42's research found hallucinated non-existent domain (NXD) rates of 44.6% for one model family and 27.5% for another across queries about 913 well-known global brands in technology, finance, healthcare, e-commerce, government, logistics, and gambling [1]. Temperature configuration influenced the NXD rate — creative settings produced 43.1% NXD compared to 32.5% at balanced settings — but no configuration eliminated the problem [1].

The hallucination rate persisted across all temperature configurations tested. Whether architectural changes, retrieval augmentation, or output grounding can substantially reduce this rate at scale remains an active area of research and should not be assumed solved by prompt-level tuning alone.

The Registration Window

The critical enabler for phantom squatting is the gap between when a model reliably generates a hallucinated domain and when an adversary registers it. Unit 42 coined the term Adversarial Exploitation Window (AEW) to describe this interval. Across multiple real-world cases their monitoring pipeline captured, the AEW ranged from 18 to 51 days [1]. This interval may exist in part because adversaries face

a resource constraint: they cannot register every hallucinated domain simultaneously. Instead, they appear to prioritize high-traffic brands and domains with high Thermal Hallucination Persistence (THP) — those generated consistently across multiple prompts and model configurations, indicating the model treats the URL as near-factual.

From the defender's perspective, this window is an opportunity. A hallucinated domain that has not yet been registered is a potential threat artifact that can be monitored and potentially preemptively claimed or blocked before it becomes live attack infrastructure.

Security Analysis

The Montana Empire Incident

The most fully documented phantom squatting case in the Unit 42 dataset is the Montana Empire phishing kit, which Unit 42 tracked from prediction to active weaponization. On March 8, 2026, Unit 42's hallucination-monitoring pipeline identified a domain resembling the e-commerce marketplace of a national postal service as a high-persistence hallucination target. Twenty-three days later, on March 31, an adversary registered that exact domain and stood up a fully functional phishing operation [1][5].

The infrastructure the attacker deployed was distinctive for two reasons. The phishing kit featured a real-time storefront scraper that mirrored the legitimate postal service's marketplace in near-real-time, a PHP backend for credential capture, and a Telegram-based command-and-control channel that allowed the operator to manually approve one-time passcodes during active sessions. The kit harvested payment card numbers, bank transfer details, and national identity document data. Forensic analysis of the attacker's development environment indicated they used an AI coding assistant to build significant portions of the kit itself — the same class of technology that generated the hallucinated domain was leveraged to build the infrastructure for exploiting it [1].

Unit 42 also validated a second case involving historical data. Hallucination monitoring had flagged a domain linked to a major UAE bank before an adversary was observed using it for credential harvesting — demonstrating that model-generated hallucinations for well-established brands can persist and remain exploitable on extended timescales, in this instance approximately eleven months [1].

Anatomy of a Phantom Squatting Attack

The phantom squatting lifecycle follows four distinct phases that defenders must understand to disrupt it effectively. In the discovery phase, adversaries deliberately probe one or more LLM systems with questions about brand websites, API endpoints, developer portals, and download locations. They collect the hallucinated outputs, de-duplicate them, and score them by hallucination persistence — the more consistently and confidently a model generates a given domain across varied prompts, the more valuable it is as a registration target.

In the registration phase, the adversary claims unregistered phantom domains, prioritizing those with high persistence scores and high-value brand associations. Domain registration typically costs $10–$20 per year, a negligible investment relative to the potential value of harvested credentials — a single convincing phishing domain targeting customers of a major financial institution or postal carrier can harvest significant volumes of credentials before defenders respond.

The lure phase operates without any active adversary involvement. When an LLM user — or an autonomous AI agent — asks the model for a URL associated with the impersonated brand, the model generates the hallucinated address and presents it as authoritative. The user or agent visits the now-registered phishing site, which may present a convincing brand clone. There is no phishing email, no malicious advertisement, and no search engine to filter; the LLM itself has delivered the victim to the attacker's infrastructure.

Finally, the bypass phase reflects a structural advantage that makes phantom squatting particularly difficult to detect with conventional defenses. A newly registered phantom domain has no threat intelligence history. It carries no reputation score. It has not appeared on any blocklist. At the moment traffic first arrives, it is indistinguishable from a newly registered legitimate domain [1]. By the time threat intelligence feeds have accumulated enough signal to flag it, victims may have already been compromised.

Amplification in Agentic AI Systems

The threat posed by phantom squatting scales with the autonomy of the AI systems exposed to it. A human user who receives a hallucinated URL from a chatbot must still choose to visit that URL, introducing a potential verification step. An autonomous AI agent that receives a URL as part of its task context — for instance, an agent retrieving API documentation, checking a software dependency registry, or automating a web-based workflow — may fetch and act on that URL without any human decision point in the loop.

Unit 42's research explicitly identified agentic deployments as the highest-consequence target for phantom squatting [1]. In an agentic context, compromise does not necessarily require credential theft. A malicious response from a phantom domain could inject adversarial instructions into the agent's context window, exfiltrate environment variables and authentication tokens, or poison a software build pipeline by substituting a malicious dependency for an expected one. The agent's own autonomy becomes the attack's delivery mechanism.

This risk is already materializing. A July 2026 analysis found that AI coding agents routinely skip package verification steps when resolving dependencies, treating LLM-recommended package sources as authoritative without independently confirming they resolve to expected registries [4]. The same trust pattern that enables phantom squatting of web domains applies directly to agentic package resolution workflows.

Hallucination Breakdown by Type

Not all hallucinated URLs present equivalent risk. Unit 42's analysis found that 49.7% of hallucinated URL errors occurred at the path level — the model invented a path or endpoint beneath an otherwise valid domain. Subdomain-level hallucinations accounted for 39.5%, and pure domain-level hallucinations (inventing an entirely nonexistent registerable domain) accounted for 10.8% [1]. The two LLM families studied showed different distribution patterns: one model skewed heavily toward path-level hallucinations at 56.6%, while the other generated pure domain hallucinations at twice the average rate, at 20%.

This breakdown matters for defenders. Path-level and subdomain-level hallucinations may not create registerable domains that adversaries can claim, but they represent traffic misdirection risks that can enable credential interception at the DNS or HTTP layer. Pure domain hallucinations represent the highest phantom squatting risk because they create entirely new registerable infrastructure under adversary control.

Sector Exposure

The Unit 42 dataset covered seven sectors. Finance and e-commerce organizations represent particularly high-value targets because hallucinated domains for those brands are likely to attract users seeking to complete high-stakes transactions — payments, fund transfers, account access — that yield credential data with immediate monetary value. Healthcare organizations present a distinct risk profile because the data harvested — health records and identity documents — carries high value for insurance fraud and identity theft, and healthcare portals frequently handle sensitive transactions that users expect to complete online.

Government services present a substantial and incompletely addressed exposure. Citizens increasingly use AI assistants to navigate government websites, locate forms, and identify contact information. A hallucinated government domain that an adversary registers and populates with a convincing agency clone can harvest identity documents and tax information from users who have every reason to trust a recommendation from an AI assistant they consider authoritative.

Recommendations

Immediate Actions

Organizations should begin auditing how their own brand appears in LLM outputs. This means systematically querying multiple publicly available LLMs for official website URLs, API endpoints, developer portals, customer login pages, and download locations associated with the organization's products and services. Any URL generated by a model that does not correspond to a registered, legitimately operated domain is a candidate phantom squatting target that should be evaluated for preemptive registration or DNS monitoring.

Security teams should extend their existing domain monitoring programs to include hallucinated variants. Most enterprise domain monitoring services support alert-on-registration for defined keyword patterns; organizations should populate those watchlists with the hallucinated domains identified through LLM auditing. The 18-to-51-day adversarial exploitation window documented by Unit 42 is enough time to detect, assess, and respond to a registration event before significant victim traffic accumulates [1].

Where budget permits, organizations should consider preemptively registering the most persistent hallucinated variants of their own domains — particularly those involving subdomains or path-level variants on primary domains — and pointing them to an explanatory landing page or redirect to the legitimate site. This eliminates the adversary's ability to claim those specific assets.

Short-Term Mitigations

For organizations deploying agentic AI systems, URL verification should be treated as a mandatory capability rather than an optional enhancement. Agents that fetch external URLs as part of automated workflows should validate those URLs against an authoritative allowlist before issuing requests. Where an allowlist is not feasible, agents should be configured to reject URLs generated entirely from LLM context and require human confirmation before fetching resources from newly encountered domains.

Development teams integrating AI coding assistants into build pipelines should implement package verification checks that confirm every recommended dependency resolves to an expected, known-legitimate registry entry before installation. This is documented practice for typosquatting defense and applies with equal force to the slopsquatting and phantom squatting variants. Dependency pinning and lockfile enforcement are foundational mitigations [2][3].

Organizations should incorporate hallucination-awareness training into security awareness programs for any staff who regularly use AI assistants. Users should understand that LLM-generated URLs are not verified references and should be independently confirmed via browser search or bookmarked links before entering credentials. This guidance is particularly important for helpdesk, customer-facing, and financial operations staff who may use AI assistants to look up service portal URLs.

DNS resolver controls offer a low-friction layer of protection. Organizations running internal DNS resolvers can subscribe to threat intelligence feeds that include phantom squatting domains as they are identified, enabling automated blocking before internal users or agents reach the malicious infrastructure.

Strategic Considerations

The phantom squatting threat vector is expected to grow in proportion to enterprise adoption of LLMs and agentic AI systems. As AI assistants become an increasingly common way employees and customers navigate organizational services, the sophistication of phantom squatting operations is expected to increase proportionally. The Montana Empire incident — in which the adversary used an AI coding assistant to build the phishing kit itself — illustrates the feedback loop enabled by accessible AI tooling: the same class of technology that generates hallucinated domains can accelerate the construction of infrastructure to exploit them. Whether AI-assisted offense is advancing faster than AI-aware defense at the industry level remains an open empirical question.

Vendors of large language models should be expected to treat hallucination-driven domain generation as a product safety issue. URL output grounding — validating model-generated domain outputs against live DNS before presenting them to users — is an approach that LLM vendors should evaluate and invest in developing as a standard capability for products providing URL recommendations. Organizations procuring AI systems should include hallucination grounding for URL outputs as a vendor evaluation criterion in their AI procurement assessments.

Industry-wide coordination on hallucinated domain sharing would substantially improve collective defense. A shared registry of model-generated phantom domains — analogous to the coordinated disclosure mechanisms used for CVEs — would allow organizations, registrars, and threat intelligence

providers to act on hallucination data before adversaries do. CSA's AI Controls Matrix and STAR for AI programs provide natural governance structures for this kind of cross-organizational intelligence sharing.

CSA Resource Alignment

The phantom squatting threat intersects directly with several CSA frameworks and initiatives. CSA's MAESTRO (Multi-Agent Extensible Security Threat Research Operations) framework addresses concerns for agentic AI deployments that encompass URL resolution integrity, particularly in the context of tool call authentication and the validation of externally sourced resources that agents fetch autonomously. The phantom squatting attack lifecycle — where an agent trusts an LLM-generated URL without verification — maps to MAESTRO's threat category of manipulated external resource access and underscores the need for agent-level URL allowlisting and provenance checks.

The AI Controls Matrix (AICM) v1.0 addresses the supply chain dimensions of this threat across multiple control domains. Controls governing AI system component provenance and integrity in the AICM AI-SC domain are directly applicable to the package-level variant of phantom squatting documented in slopsquatting research. Controls addressing vulnerability monitoring for AI-integrated systems in the AICM AI-VM domain extend naturally to monitoring for hallucinated domain registrations as an AI-specific vulnerability class. Organizations using the AI-CAIQ for supplier assessments should add questions about hallucination grounding for URL outputs when evaluating LLM providers and AI coding assistant vendors.

CSA's STAR for AI program provides the attestation and assessment infrastructure to operationalize these controls at scale. An organization seeking third-party validation of its phantom squatting defenses — including LLM output monitoring, domain watch programs, and agent-level URL verification — can document those controls through a STAR Level 1 self-assessment or request a Level 2 third-party audit.

The Zero Trust guidance published by CSA also applies here, specifically the principle that no resource — including a URL generated by a trusted AI system — should be considered implicitly trusted without verification. Phantom squatting is precisely the attack that Zero Trust principles are designed to mitigate: it exploits implicit trust in AI-generated content to bypass perimeter controls entirely.

References

[1] Keerthiraj Nagaraj, Diva-Oriane Marty, Beliz Kaleli, Oleksii Starov. "Phantom Squatting: AI-Hallucinated Domains as a Software Supply Chain Vector." Palo Alto Networks Unit 42, June 30, 2026.

[2] Help Net Security. "Package hallucination: LLMs may deliver malicious code to careless devs." Help Net Security, April 14, 2025.

[3] Cloud Security Alliance AI Safety Initiative. "Slopsquatting: AI Code Hallucinations Fuel Supply Chain Attacks." Cloud Security Alliance, April 19, 2026.

[4] TechTimes. "AI Coding Agents Skip Package Verification, and Attackers Are Exploiting It." TechTimes, July 1, 2026.

[5] Cyberpress. "Montana Empire Phishing Kit Abuses AI-Hallucinated Domain to Steal Credentials." Cyberpress, July 1, 2026.

[6] Trend Micro. "Slopsquatting: When AI Agents Hallucinate Malicious Packages." Trend Micro, 2025.

Further Reading

- The Hacker News (Swati Khandelwal). "Phantom Squatting Uses AI-Hallucinated Domains for Phishing and Malware." The Hacker News, July 1, 2026.
- Cybernews. "'Phantom squatting' uses AI hallucinated domains for cyber attacks." Cybernews, July 2026.
- GBHackers. "Attackers Register AI-Hallucinated Domains to Deliver Phishing Kits and Malware." GBHackers, July 1, 2026.
- Dark Reading. "'Phantom Squatting': An Emerging AI-Driven Supply Chain Threat." Dark Reading, July 2026.