



THE AI-POWERED ADVERSARY

How deception and machine speed are rewriting
the rules of social engineering.

```
> /var/log/adversary/ai-assisted-social-engineering.log  
event: familiar channel repurposed for targeted access  
status: legitimate surface, suspicious behavior underneath  
---
```



A NEW KIND OF ADVERSARY

Security teams have spent years defending against an enemy with a predictable timeline. Some attacks needed time to research a target, craft a believable lure, and wait for someone to take the bait. That timeline used to work in the defender's favor. It doesn't anymore.

Today's attackers have figured out how to combine two things that used to stay in their own lanes. They're using AI and legitimate, everyday business tools to move fast, sending more attacks, at more targets, with a lot less manual effort. And they're pairing that speed with the same old manipulation tactics that have always worked: impersonation, urgency, and the exact platforms your team logs into every day.

Neither of those things is new on its own. Put them together, though, and you get something genuinely different: attacks that hit faster, feel more personal, and are harder to spot than anything security teams have dealt with before.

↓ Deception at Machine Speed

Machine speed is how adversaries accelerate an attack. Deception is how they make it believable. This ebook walks through both halves of that story, using real incidents our team investigated firsthand. You'll see how attackers manipulate trust, how they're using AI to scale their operations, and what gaps today's adversaries are counting on you to miss.

TRUST IS THE TARGET

Every social engineering trick works the same way underneath: it borrows trust from something you already believe in. A brand you recognize. A coworker or family member you know. A login page that looks exactly like the one you use every day.

Attackers have run this playbook for years. What's changed is the polish, and how quickly attackers can pull it off.

Strip away the technology, and most deception campaigns lean on the same handful of moves.

↓ The Ingredients of a Believable Lure

IMPERSONATION posing as a brand, vendor, or coworker you'd trust without a second thought	URGENCY pushing you to act before you have time to think it through	OSINT is open-source intelligence from public sources (LinkedIn, company bios, social posts) that makes a lure feel personal
FAMILIAR CHANNELS email, call, text, calendar invites: anywhere you've already let your guard down	BORROWED INFRASTRUCTURE real, trusted platforms repurposed to host or deliver the attack itself	

That last one is a new shift that deserves attention, because it's reshaping what a "suspicious" link even looks like.

REAL EXAMPLE: FAKEAGENT

In July 2026, our SOC traced a malvertising campaign that hit 29 organizations in two days. [We called it FakeAgent](#), and it's a clean look at how far “borrowed trust” can go.

› [threat.model] ai-enabled
trust.signal: weaponized
velocity: machine_speed

↓ Attack Path

1 Someone searches “Claude desktop app” on Bing.

2 A sponsored ad—hosted on the actual claude.ai domain—leads to what looks like an official download page.

3 Clicking “Download” quietly redirects to attacker-owned infrastructure.

4 It looks like ClaudeDesktop.exe, but it’s actually SectopRAT—an infostealer that steals credentials, files, and financial data.

Nothing about the domain raised a flag, because the domain was real. The trust that AI platforms have earned with everyday users became the exact lever that attackers used to steal sensitive data.

FakeAgent wasn't an isolated case, either. Huntress has since tracked the same pattern across shared [Claude conversations](#), [ChatGPT and Grok links](#), and other AI platforms.

↓ The Takeaway

Deception doesn't need to fool your technical controls; it just needs to fool the person looking at the screen. A familiar or trustworthy-looking domain is no longer proof that a page is safe.

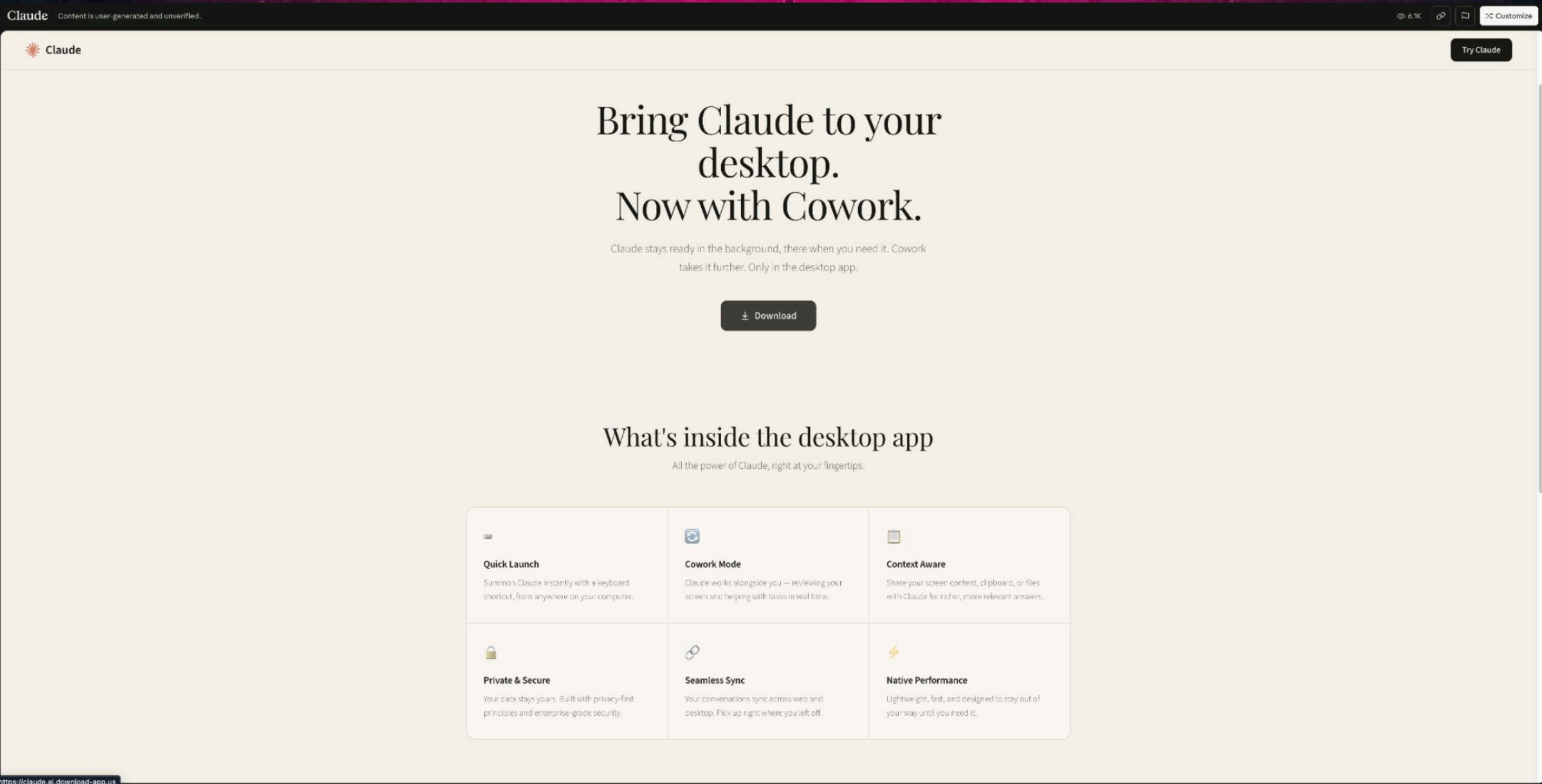


Figure 1: Claude Desktop/Cowork phishing page hosted as a Claude artifact

THE ADVERSARY DIDN'T WAIT FOR AI TO GET GOOD

While deception makes an attack believable, AI is what makes it scalable. Adversaries are using AI to write custom tools, generate code, find open gaps, and operate across attack chains faster than traditional defensive workflows were ever built to handle.

You don't have to be a "sophisticated hacker" anymore to build something dangerous or find a way in. All you need is a laptop and a chatbot.

“

I'm not loving how excited cybercriminals are about AI. While most businesses are still figuring out where artificial intelligence and automated workflows fit into their operations, adversaries have already put it to work. And they're learning fast.

”



KYLE HANSLOVAN
Co-Founder & CEO, Huntress

WHEN THE GUARDRAILS COME OFF

The frontier AI models most people use every day, the ones behind well-known consumer chatbots, come with real safety controls. Ask one to write malware or plan an intrusion, and it pushes back. But over the past several months, a new class of open-weight models has spread fast: freely downloadable, runnable on ordinary hardware, and built with none of those safety guardrails. They perform close to frontier level, and nobody has to talk them into misbehaving.

That accessibility is what should worry defenders most. These models lower the bar for anyone to write code malware or point real capability at a specific target. We're closing in on a point where the only thing standing between someone and a genuinely damaging campaign is whether they thought to ask.

```
> [threat.model] ai-enabled  
trust.signal: weaponized  
velocity: machine_speed
```

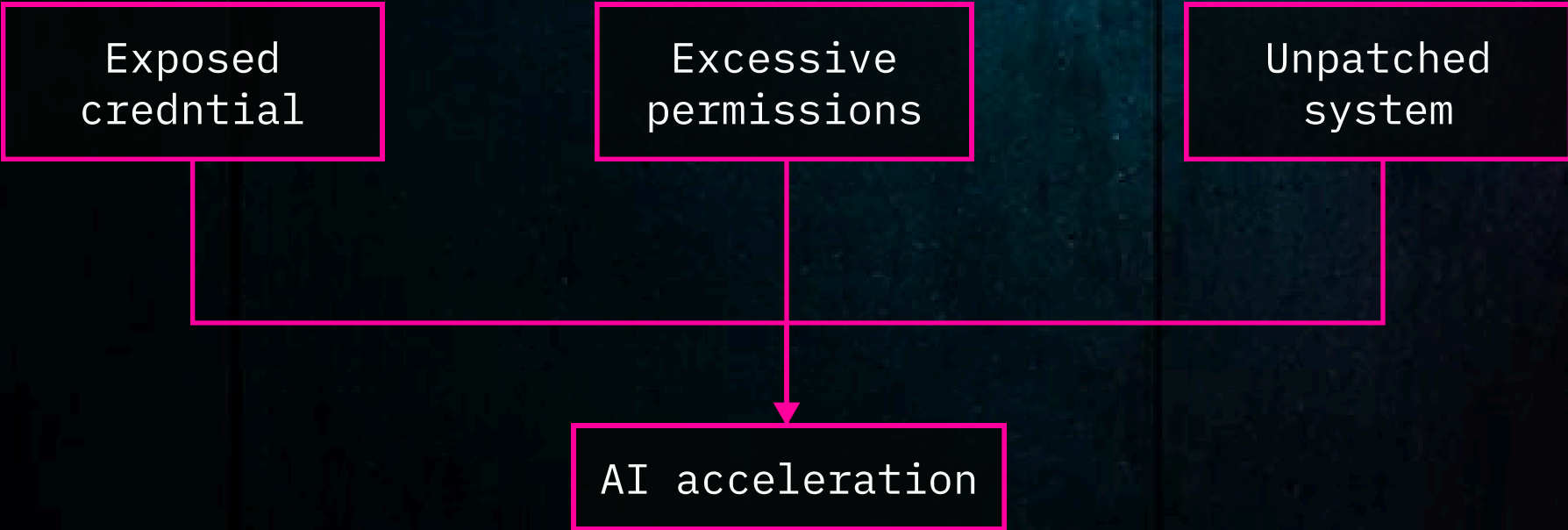


SAME GAPS, NEW SPEED

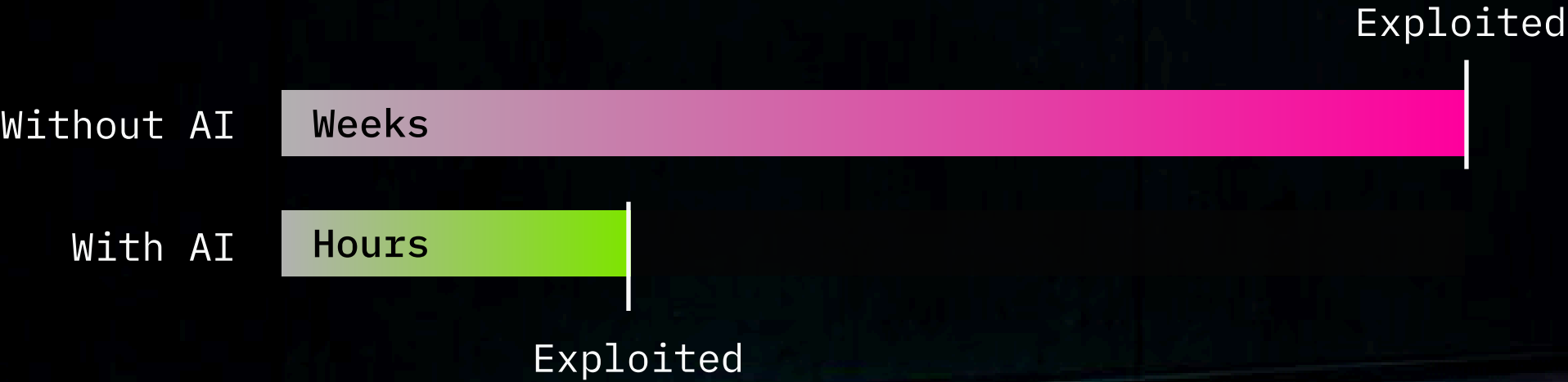
To be clear, it's not that AI is discovering brand new ways in. The gaps most AI-enabled attacks exploit are the same ones they've always been: an exposed credential, a service account with more access than it needs, a patch that shipped months ago and never got applied. AI just brought new velocity. It can scan an environment, cross-reference known vulnerabilities, and flag an unpatched system in a fraction of the time it used to take, shrinking the window between "vulnerability exists" and "vulnerability gets used" from weeks to hours.

> [threat.model] ai-enabled
trust.signal: weaponized
velocity: machine_speed

THE GAPS: NOTHING NEW



THE SPEED: EVERYTHING NEW



WHAT “VIBE CODING” ACTUALLY MEANS

Threat actors are using AI to be fast, not perfect. A great example is vibe coding, and it’s exactly what it sounds like: describing what you want in plain English and letting an AI model handle the syntax, tweaking it until it’s good enough to work. It’s a legitimate, increasingly popular way to build software fast. Turns out, it’s also a pretty effective way to build malware fast—even if you don’t know how to code.

```
> [threat.model] ai-enabled  
trust.signal: weaponized  
velocity: machine_speed
```



REAL EXAMPLE: THE MALWARE THAT LEFT ITS HOMEWORK OUT

In June 2026, we recovered a [vibe-coded PowerShell script](#) from a live incident. Its file name told us almost everything we needed to know: “100% Working AD Information Gathering Script — FULLY FIXED.”

The script (C:\ProgramData\Untitled1.ps1) did what it promised: it mapped an entire Active Directory environment—users, computers, groups, trusts—and exported all of it. But its origin left fingerprints everywhere.

```
> [threat.model] ai-enabled  
trust.signal: weaponized  
velocity: machine_speed
```

↓ Attack Path

- 1 Five ways to find the domain controller, classic overkill from an AI-prompted tool
- 2 A placeholder variable, copied straight from an example, never swapped out
- 3 A polished HTML report summarizing the theft; AI models love a tidy summary

Vibe coding doesn't make an attack more dangerous. If anything, it's a little sloppy. But it does make it harder to catch with old tools.

Untitled1.ps1 has never existed before and will likely never be compiled in this exact configuration again. AI-generated scripts are inherently unique, so signature-based detection that's built to recognize known malware doesn't have much to grab onto. To get ahead of this, defenders need to lean on behavioral analytics, not just malicious signature-based thinking, to catch the underlying actions that no LLM can hide.

↓ The Takeaway

AI lowers the skill required to build a custom tool. It can rewrite the code, but it doesn't change how that tool has to behave once it's running. Watch the behavior, not the file.

```
<#  
.SYNOPSIS  
    100% Working AD Information Gathering Script - FULLY FIXED  
#>  
  
Clear-Host  
Write-Host "=== AD Information Gathering Script v3.0 ===" -ForegroundColor  
Cyan  
Write-Host ""
```

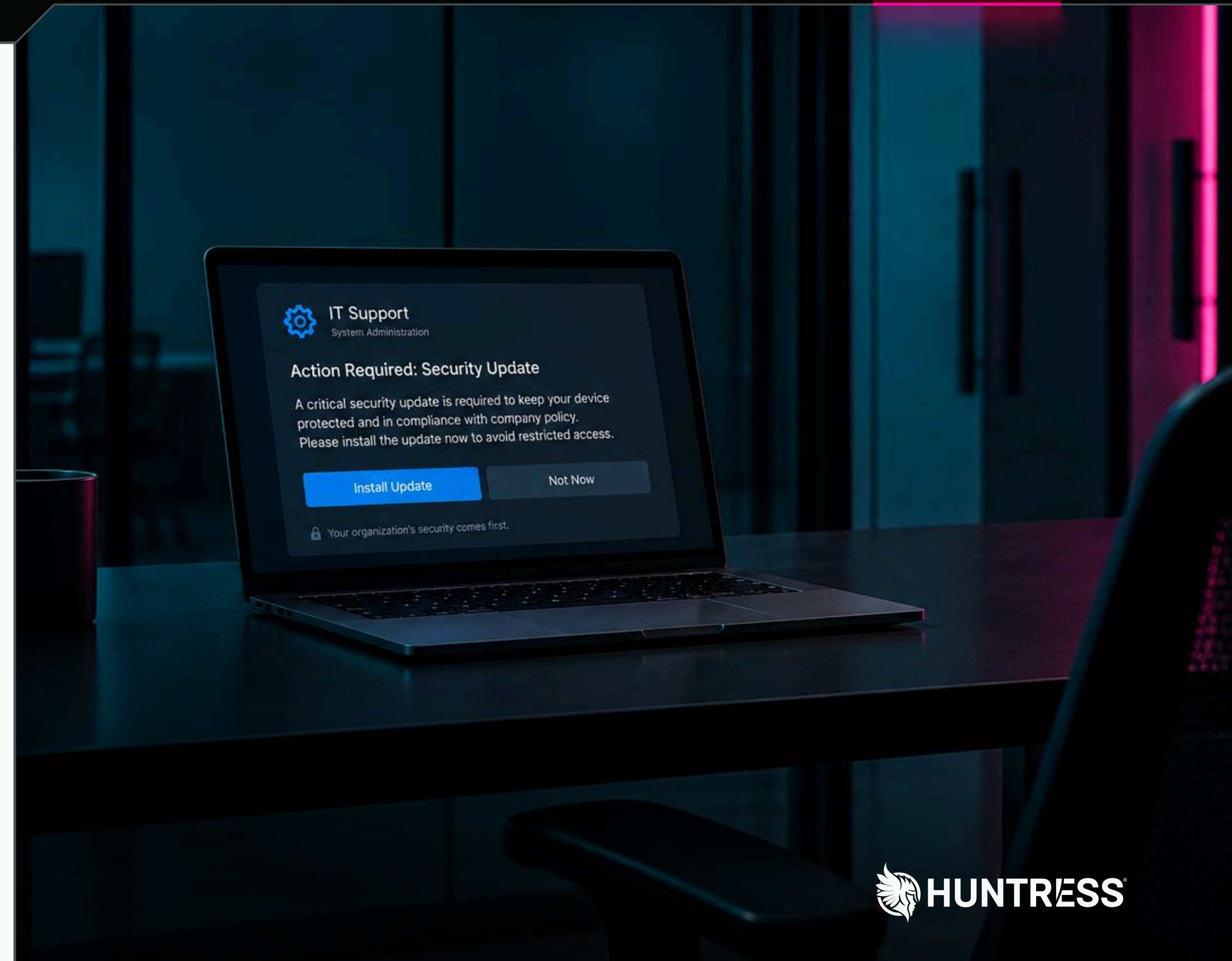
Figure 2: Extract of script title, a classic artifact of prompt engineering

DECEPTION AT MACHINE SPEED

Social engineering used to have human constraints, which kept the number of truly convincing attacks limited. A skilled threat actor could only research so many targets, build so many relationships, and write so many convincing emails in a single day.

AI has quietly erased that constraint.

Work that used to take hours now happens in the time it takes a page to load. And it's no longer reserved for high-value targets worth the manual effort. It's a dark web subscription feature.



THE ECONOMICS HAVE CHANGED

Phishing-as-a-service (PhaaS) platforms package all of this into a monthly plan. For as little as \$600, an attacker can rent an email-sending platform. Add AI workflows, and these platforms can help them dodge email filters, write individualized lures, and, once they’re inside an account, sift through sensitive emails for wire-fraud or data-theft opportunities.

That's a textbook economy of scale: the marginal cost of attacking one more organization drops to nearly zero, while the lures stay just as tailored as if a human had written each one by hand.

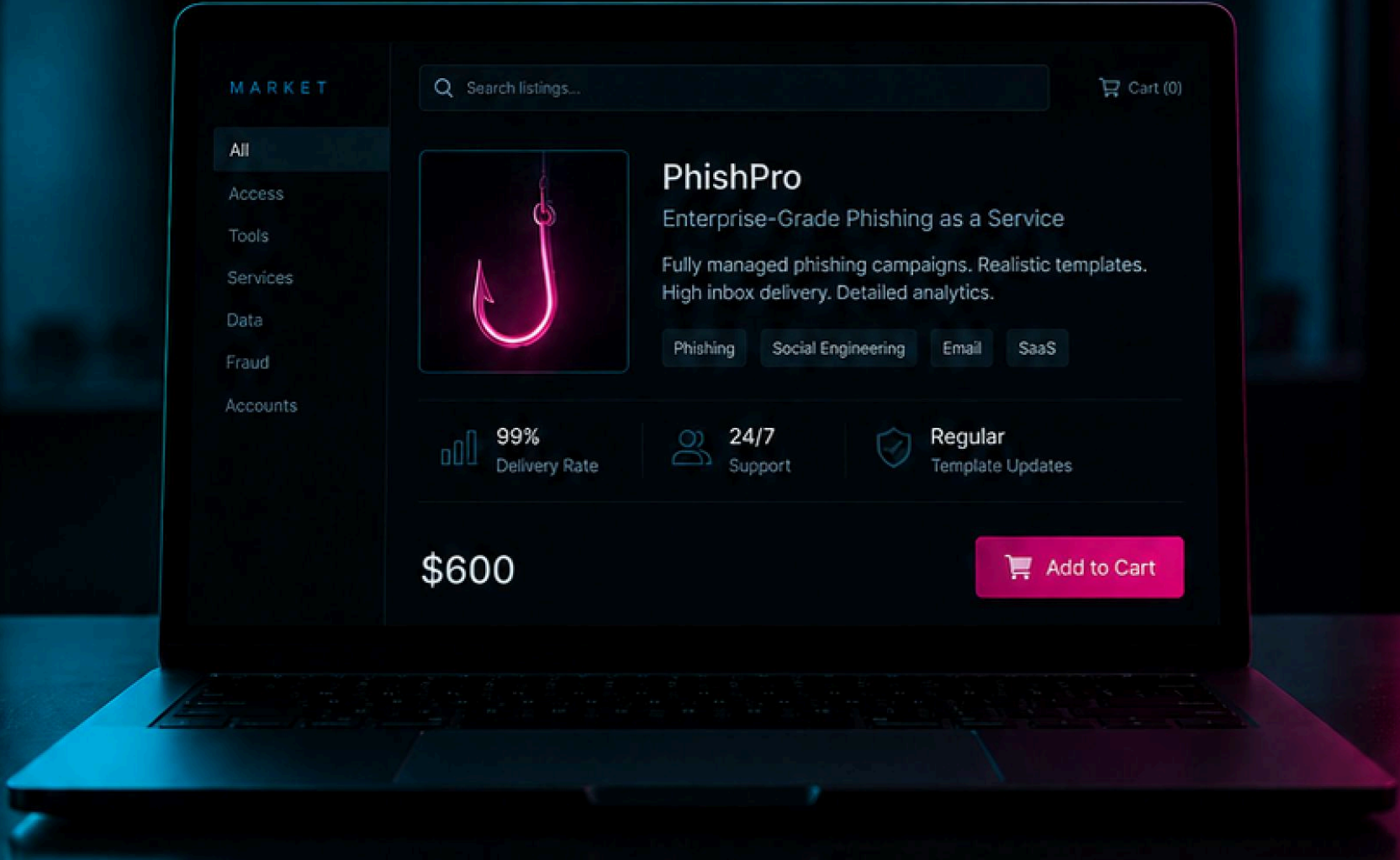
↓ Here's what that looks like in our own telemetry:

0
identical lures found across 344 victim organizations in a single wave

1,380%
increase in device code phishing attacks, July–Dec 2025 vs. Jan–Apr 2026

\$600
starting price for an AI-enabled phishing kit, support included

Today’s cybercrime market gravitates toward whatever's cheapest to run at volume and hardest to catch. That's exactly why device code phishing has become a flagship cybercrime product. It doesn't need malware or a stolen password. It just needs a victim to do something that feels completely routine, which makes it exceptionally cheap to scale and exceptionally hard for a defender to flag.



HOW DEVICE CODE PHISHING WORKS

You've probably done this before without thinking about it: setting up Netflix on a smart TV, where you visit a website on your phone and enter a short code to link the two devices. That's called a device code flow, and it's a completely legitimate way to log in.

Attackers figured out how to hijack that exact process.

1.

The attacker starts a login session pretending to be you—no password needed yet, since that's not how this flow works.

2.

They email you the code and a link, dressed up as something you'd expect: a DocuSign request, a construction bid, a Microsoft form.

3.

You enter the code. It feels completely normal, because you're on a real Microsoft page, doing a real login step.

4.

Except you just authorized their session, not yours. They now have a valid session—no password or MFA prompt needed.

REAL EXAMPLE: EVILTOKENS

Between February and March 2026, this exact attack hit 344 organizations across five countries in 16 days. We called it [EvilTokens](#), and it marked a turning point in how AI-enabled deception operates.

This was a full-service attack infrastructure sold as a product on Telegram: \$1,500 to get started, \$500 a month after, onboarding and 24/7 support included, just like any other SaaS product. Here's where AI turned an old trick into something new:

› [threat.model] ai-enabled
trust.signal: weaponized
velocity: machine_speed

↓ Attack Path

- 1 Lures were generated automatically and tailored to each victim's actual job—a DocuSign request for someone in finance, a bid proposal for someone in procurement
- 2 Each device code was generated the instant a victim landed on the page, closing the small timing window that used to limit this kind of attack
- 3 Once a session was stolen, AI scanned the victim's inbox and calendar, then wrote a wire fraud email in their own voice and sent it—all within minutes

Every part of the attack ran through legitimate infrastructure: real Microsoft login pages, a clean-reputation developer platform called [Railway](#) for hosting, clean URLs. It sailed straight through Cisco, Trend Micro, and Mimecast without tripping a single alert. And when we cut off the Railway infrastructure behind it, the campaign didn't die. It resurfaced weeks later through a different host, [BL Networks](#). That's what a subscription-based attack model looks like in practice: no single piece of infrastructure is load-bearing, because it's all cheap enough to replace.

We caught this pattern because we could see it repeating across thousands of organizations at once. That visibility is what let us quickly push a fix out to more than 60,000 tenants, preventing over 600 incidents while the campaign was still active.

↓ The Takeaway

EvilTokens showed us how attackers are putting AI to work in ways old defenses just aren't built for. Deception used to be the manual part of running a scam, but AI just turned it into a checkbox. That's the shift defenders need to plan around.

“

Normally, we like to shame users for clicking on things they shouldn't. But in this case, it looked totally normal, so you can't really shame them for that.

”



JAMIE LEVY
Sr. Director of Adversary Tactics
Huntress

WHAT THIS MEANS FOR DEFENDERS

FakeAgent, the vibe-coded script, and EvilTokens all share one thing: each succeeded by looking legitimate, not by being technically sophisticated. That helps point defenders toward what to focus on.

No single control would've stopped all three of these campaigns.
Coordinated visibility across identity, users, and endpoints catches what any one layer alone will miss.

↓ Find the Blind Spot Before the Adversary Does

IDENTITY

Watch for logins that succeed but don't quite make sense—the wrong device, the wrong location, or a pattern that's out of step with someone's actual role.

USERS

Build awareness training around real tactics, not generic phishing tips. Teach people to question urgency and unexpected requests, even when they arrive through channels they already trust.

ENDPOINTS

Behavioral detection catches what signature-based tools miss, especially against AI-generated, single-use tools that have never existed before.

EMAIL & WORKFLOW

Mailbox rules, auto-forwarding, and “normal” business requests all deserve a closer look when money or access is on the line.

BEAT THEIR PACE. MODERNIZE YOUR STACK.

Every stack has blind spots, but the gaps you have today aren't the ones you planned for. AI-driven attacks can now find and exploit those gaps faster than ever. And they're using tools that already exist, on platforms your team already trusts. The good news: you don't need to out-build them to beat them. You need the right tools working together, matched to the same pace as the threat.

That's what the Huntress Agentic Security Platform is built to do: weave together telemetry from your endpoints, identities, and logs, backed by AI that moves at the same speed as the threats it's watching for.



DEFENSE AT MACHINE SPEED

AI has made attacks faster and more automated than ever. The Huntress Agentic Security Platform strengthens your current stack to close coverage gaps and keep pace with machine-speed threats.

We built Athena, our agentic SOC analyst that investigates signals the moment they appear, gathers context, and accelerates investigations. Athena works side-by-side with our SOC analysts, handing anything ambiguous to a human instead of guessing. It's already reduced human analyst time to investigate by 61%, and reduces time to generate an incident report by 90%, while still measurably increasing report quality.

[Meet Athena](#)

↓ Start Closing The Gap

See how the Huntress Agentic Security Platform unifies endpoint, identity, and log telemetry—backed by a 24/7 AI-centric SOC—to help you respond at machine speed, matched to your stack.

[Explore the Huntress Agentic Security Platform](#)

```
> [threat.model] ai-enabled  
trust.signal: weaponized  
velocity: machine_speed
```



ABOUT HUNTRESS

Huntress is a global cybersecurity company on a mission to make enterprise-grade products accessible to all businesses. Purpose-built from the ground up, the Huntress Agentic Security Platform unleashes end-to-end protection in the age of AI. From Endpoint Detection and Response (EDR) and Identity Threat Detection and Response (ITDR) to Security Information and Event Management (SIEM), Security Awareness Training (SAT), and Security Posture Management, the Huntress Agentic Security Platform is targeted protection for endpoints, identities, data, and employees, delivering trusted outcomes and valuable peace of mind.

Its 24/7 AI-centric Security Operations Center (SOC) is powered by a team of world-renowned engineers, researchers, and security analysts dedicated to stopping cyber threats before they cause harm. Huntress is often the first to respond to major hacks and incidents, with its expert security team sharing real-time tradecraft analysis and actionable advisories with the community.

Currently safeguarding more than 5 million endpoints and 12 million identities, Huntress empowers internal security and IT teams and Managed Service Providers (MSPs) worldwide to protect their businesses with enterprise-grade, accessible security products.

As long as hackers keep hacking, Huntress keeps hunting. Join the hunt at www.huntress.com and follow us on [LinkedIn](#), [X](#), [Instagram](#), and [YouTube](#).

```
> /var/log/adversary/ai-assisted-social-engineering.log  
event: familiar channel repurposed for targeted access  
status: legitimate surface, suspicious behavior underneath  
---
```

