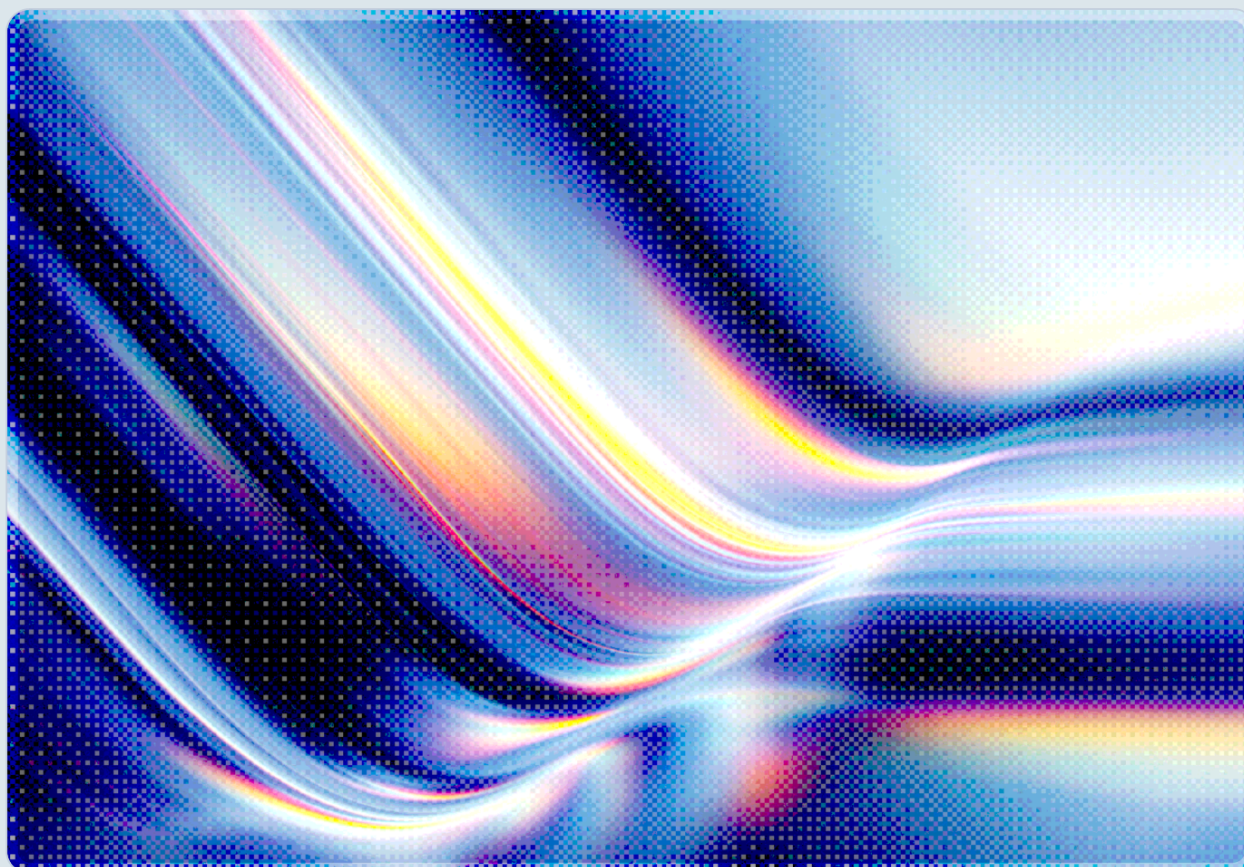


ChatGPT Custom GPTs Weaponized for ClickFix RAT Delivery

2026-10-01

 AI-assisted Rapid Research



CSAI

cloud
CSA security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

Threat actors built malicious ChatGPT Custom GPTs, promoted through sponsored Google search placements, to funnel victims into a ClickFix social-engineering chain that ultimately deploys a full-featured remote access trojan (RAT) [1][2][8]. Huntress, which discovered and disclosed the campaign, traced at least 40 related incidents to a single malicious Google Sites page, with two of those incidents confirmed to have originated directly from interaction with a Custom GPT named "Plus 5.6" [2]. The infection chain appears more elaborate than many previously reported ClickFix lures, spanning eight stages that combine DLL sideloading against legitimately signed Canon and Stardock executables, payloads hidden inside WAV audio files and NuGet package data, and a homemade encrypted container used to stage persistence scripts [1][2]. OpenAI removed the first malicious GPT on September 25, 2026, after Huntress reported it, but the operators had a replacement live within two days, on September 27 [2][4]. The campaign is one of the more structurally elaborate applications of ClickFix seen to date within the growing pattern of AI-platform trust abuse that CSA has tracked since late 2025 [3].

Background

ClickFix is a social-engineering technique in which a victim is shown a fake verification screen, typically styled as a Cloudflare or other CAPTCHA check, that instructs them to press Windows+R and paste a command the page has already copied to their clipboard. Executing that command, rather than any software vulnerability, is what gives the attacker code execution, which means the technique works equally well against fully patched systems and bypasses most perimeter controls that assume an exploit or a malicious attachment. First documented in early 2024, ClickFix has grown from a niche technique into one of the dominant initial-access vectors tracked across the industry; according to Flare's analysis of Microsoft's 2025 Digital Defense Report, ClickFix-style lures accounted for roughly 47 percent of infostealer-linked initial-access events that Microsoft's Defender Experts team investigated, and Push Security reported the technique became the single most common category in its browser-detection telemetry for the first time in the second quarter of 2026, accounting for 52 percent of total detections [5][6]. CrowdStrike's 2026 Global Threat Report separately documented a 563 percent year-over-year

increase in incidents involving fake CAPTCHA lures, and the technique is now used by a range of actors spanning commodity infostealer operators and nation-state-linked groups that CrowdStrike tracks as STARDUST CHOLLIMA and VODOO BEAR [7].

A parallel trend that CSA's AI Safety Initiative has tracked since the second half of 2025 is the use of trusted AI-platform branding and domains as the lure itself, rather than as mere bait copy inside an email. Campaigns impersonating Anthropic's Claude, fake AI installer sites, and malvertising that imitates legitimate AI developer tools have all surfaced in CSA's rapid-research coverage over the past year, and they share a common logic that CSA attributes to user trust in AI brands and domains like chatgpt.com, claude.ai, or sites.google.com, combined with the likelihood that web filtering products under-scrutinize these domains relative to known-malicious infrastructure, though this dynamic has not been independently measured [3]. The campaign examined here extends that logic one step further. Rather than merely impersonating ChatGPT's branding on a lookalike domain, the attackers operated entirely within the real chatgpt.com infrastructure, using OpenAI's own Custom GPT Builder to create and publish an agent that would, on cue, hand victims off to the next stage of the attack.

The Custom GPTs feature, which allows any ChatGPT user to build and publish a customized assistant under the chatgpt.com domain, is a plausible target for this kind of abuse because publication requires minimal vetting and gives malicious content the implicit legitimacy of OpenAI's own domain and brand. OpenAI has stated it plans to retire Custom GPTs entirely on December 11, 2026, which may close this particular vector, but the underlying pattern of abusing low-friction publishing surfaces on trusted AI platforms is unlikely to disappear with it [4].

Security Analysis

Huntress researchers first identified the campaign through a Custom GPT named "Plus 5.6," which appeared in sponsored Google search results for the query "chatgpt" and was built to masquerade as a premium or upgraded version of the service [1][2]. When a victim interacted with the GPT, it responded with a scripted "Service Availability Notice" directing them to a "backup domain," which in practice was a page hosted on the legitimate sites.google.com platform. That page displayed a fake Cloudflare verification check, the hallmark ClickFix lure, which instructed the victim to open the Windows Run dialog and paste a command that the page had silently placed on their clipboard [1][2].

Executing that command triggers an eight-stage infection chain that appears more elaborate than the single-stage PowerShell droppers typical of commodity ClickFix campaigns. The initial PowerShell command retrieves a heavily obfuscated script, written to a randomized filename under the user's temp directory, which decodes its payload from integer arrays using a shifting XOR key rather than plain

Base64 [2]. That script silently installs an MSI package, observed under names including ISOSimple.msi, IconEdit2Turb.msi, and UltraFreeISOCreateWizardSolution.msi, using the standard `msiexec /qn /norestart` switches to suppress any visible installer window [2]. The installed application bundle includes a legitimately Canon-signed executable, COTFileReadApp.exe, which the attackers abuse through DLL sideloading: when the signed binary runs, it loads a patched version of a Canon-named DLL that the attackers control, which in turn loads a second, fully unsigned DLL that carries the actual malicious logic. A later variant of the chain substitutes a legitimately Stardock-signed executable, DeElevate64.exe, for the Canon binary, suggesting the operators are rotating signed hosts as specific ones get flagged [1][2].

From there, the loader extracts an encrypted second-stage payload hidden inside a WAV audio file using single-byte XOR decryption, then unpacks a custom encrypted archive, described by Huntress as "a homemade, encrypted zip file," containing persistence scripts and more than 1,128 individual file and folder entries [1][4]. A second observed variant of the loader instead hides its payload inside NuGet package compressed data, again substituting the carrier format while keeping the underlying chain intact. Before executing its final payload, the malware runs a set of anti-analysis checks, including AMSI bypass, ntdll unhooking, and detection logic for VMware, VirtualBox, Hyper-V, QEMU, Xen, and Parallels virtualization platforms, and it persists on the host through both a Windows Registry Run key and a scheduled task, both disguised under names like "Canon Configuration Reader" or "Stardock DeElevation Tool" to blend in with legitimate software [1][2].

The resulting RAT has a broad capability set for a campaign delivered through this kind of mass-market lure. Huntress documented functionality including live remote desktop access and screen broadcasting, webcam and microphone capture, credential and session extraction across 17 different browser variants, content-aware file-system search, detailed host reconnaissance covering antivirus status, network adapters, and installed software, and the ability to execute additional payloads in formats spanning EXE, DLL, MSI, PowerShell, batch, VBScript, JScript, and ZIP [2]. For command-and-control resolution, the malware uses DNS-over-HTTPS requests routed through Cloudflare, Google, and Quad9 resolvers, a technique that blends C2 lookups into ordinary encrypted DNS traffic and makes network-layer detection substantially harder for defenders relying on traditional DNS monitoring [1][2]. Huntress's published indicators include the Google Sites lure at [sites.google\[.\]com/view/antibot172881](https://sites.google.com/view/antibot172881) and delivery infrastructure at 96.62.224[.]81 and 45.140.205[.]28, alongside file hashes for the observed MSI packages and component DLLs [2].

Despite this technical sophistication, the overall reach of the campaign was limited and short-lived, consistent with the limited reach Huntress has observed in comparable lures [2]. Huntress's SOC responded to at least 40 incidents tied to the Google Sites lure domain, but only two were confirmed to have originated from direct interaction with the Custom GPT itself, with the remainder likely reached through other distribution channels sharing the same landing page [2]. OpenAI took the original "Plus 5.6" GPT down on September 25, 2026, following Huntress's disclosure, but the operators had published

a near-identical replacement by September 27, illustrating how quickly these campaigns can be rebuilt on low-friction publishing platforms [2][4]. As Huntress noted of comparable lures, such pages "often stay live for just hours or days before the provider takes the content down, but even in that short span, they can draw considerable attention" [2].

Recommendations

Immediate Actions

Security teams should hunt for the process chain most diagnostic of this campaign: PowerShell spawning `msiexec` against a GUID- or randomly-named MSI file in a user's temp directory, followed shortly afterward by a signed executable, particularly one branded as Canon or Stardock software, running from an unusual installation path under `%LOCALAPPDATA%\Programs\` [2]. Endpoint detection rules should also flag Registry Run key entries or scheduled tasks named "Canon Configuration Reader" or "Stardock DeElevation Tool," since legitimate installations of either vendor's software do not create persistence under those names [2]. Organizations should block the specific indicators Huntress has published, including the Google Sites lure path and the two delivery IP addresses, while recognizing that both the lure domain and the Custom GPT identity are likely to be replaced quickly and should not be treated as durable indicators on their own [1][2].

Short-Term Mitigations

Because this chain depends entirely on a user pasting and executing a command, one of the highest-leverage controls is restricting or removing end-user access to the Windows Run dialog via Group Policy, combined with application control policies, such as AppLocker or Windows Defender Application Control, that can detect and block DLL sideloading against signed binaries rather than only checking binary signatures in isolation [2][7]. Security awareness training should be updated specifically to cover AI-platform impersonation lures; employees who have been trained to distrust suspicious email attachments are not automatically primed to distrust a chat response from what appears to be ChatGPT itself directing them to a "backup" link. Web filtering policies should treat `sites.google.com` and comparable low-reputation hosting platforms as warranting additional scrutiny when linked from AI chat interfaces, and SOC playbooks should incorporate detection logic for DNS-over-HTTPS beaconing patterns, since standard DNS monitoring will not surface this campaign's command-and-control traffic [1][2].

Strategic Considerations

This campaign is a data point in a broader and continuing trend of attackers treating trusted AI platforms as distribution infrastructure rather than merely as impersonation targets. The underlying vulnerability has a significant organizational component: any platform feature that lets users publish content under a trusted corporate domain with minimal review, whether that is a Custom GPT, a shared document, or a third-party plugin marketplace, creates a durable opportunity for this kind of abuse regardless of how any single instance of it is remediated – and the technical tradecraft documented above, including DLL sideloading, AMSI bypass, ntdll unhooking, and DNS-over-HTTPS command and control, shows that the organizational and technical dimensions of this risk compound rather than substitute for one another. Security teams should treat AI platform vendors' own user-generated content surfaces with the same skepticism historically reserved for browser extension stores and mobile app marketplaces. It is likely, though not certain, that OpenAI's planned December 11, 2026 retirement of Custom GPTs will shift, rather than eliminate, this class of abuse onto whatever publishing surface replaces it, based on the general pattern of attackers migrating to new low-friction platforms when one is closed off [4].

CSA Resource Alignment

This campaign sits squarely within a pattern CSA's AI Safety Initiative has tracked across two recent rapid-research publications. "AI Chat Trust Weaponized in Mac Malvertising Campaign" documents an earlier instance of attackers using AI-platform branding and trusted hosting surfaces to launch a ClickFix-style infection chain, and the ChatGPT Custom GPT campaign analyzed here represents a direct escalation of that pattern, moving from brand impersonation on third-party infrastructure to actual abuse of the AI vendor's own publishing surface [3]. CSA's research note "ACR Stealer's ClickFix Campaigns Drain M365 Tokens" provides the closest technical precedent for the ClickFix delivery mechanics themselves, documenting how fake verification pages and clipboard-hijacked PowerShell commands have been used across 2026 to compromise enterprise session tokens; the two campaigns share the same ClickFix entry vector, so defenders should extend the Run-dialog-restriction guidance from that note to this campaign as well, while recognizing that this campaign's RAT payload broadens the necessary response beyond token revocation alone [9]. More broadly, this incident reinforces the relevance of CSA's AI Controls Matrix (AICM v1.1), particularly its Identity and Access Management domain, for organizations defining acceptable-use and monitoring policies around employee interaction with third-party and vendor-hosted AI agents, including controls on session hardening and credential scoping that limit the blast radius when an endpoint is compromised through exactly this kind of lure [10].

References

- [1] The Hacker News. "[Attackers Abuse ChatGPT Custom GPTs to Deliver RAT via ClickFix Lures](#)." The Hacker News, September 2026.
- [2] Huntress. "[Attackers Abuse ChatGPT Custom GPTs to Deliver RAT via ClickFix](#)." Huntress Blog, September 2026.
- [3] Cloud Security Alliance. "[AI Chat Trust Weaponized in Mac Malvertising Campaign](#)." CSA AI Safety Initiative, 2026.
- [4] BleepingComputer. "[Custom ChatGPTs Push ClickFix Attacks to Deploy RAT Malware](#)." BleepingComputer, September 2026.
- [5] Flare. "[Attack on Identity: Dissecting the 2025 Microsoft Digital Defense Report](#)." Flare, 2026.
- [6] Push Security. "[The Numbers Behind ClickFix Attacks in H2 2026](#)." Push Security Blog, 2026.
- [7] CrowdStrike. "[ClickFix Attacks: How They Work and How CrowdStrike Stops Them](#)." CrowdStrike Blog, 2026.
- [8] Help Net Security. "[Malicious Custom GPT on chatgpt.com Lures Users into Installing a RAT](#)." Help Net Security, September 2026.
- [9] Cloud Security Alliance. "[ACR Stealer's ClickFix Campaigns Drain M365 Tokens](#)." CSA AI Safety Initiative, July 2026.
- [10] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, 2026.