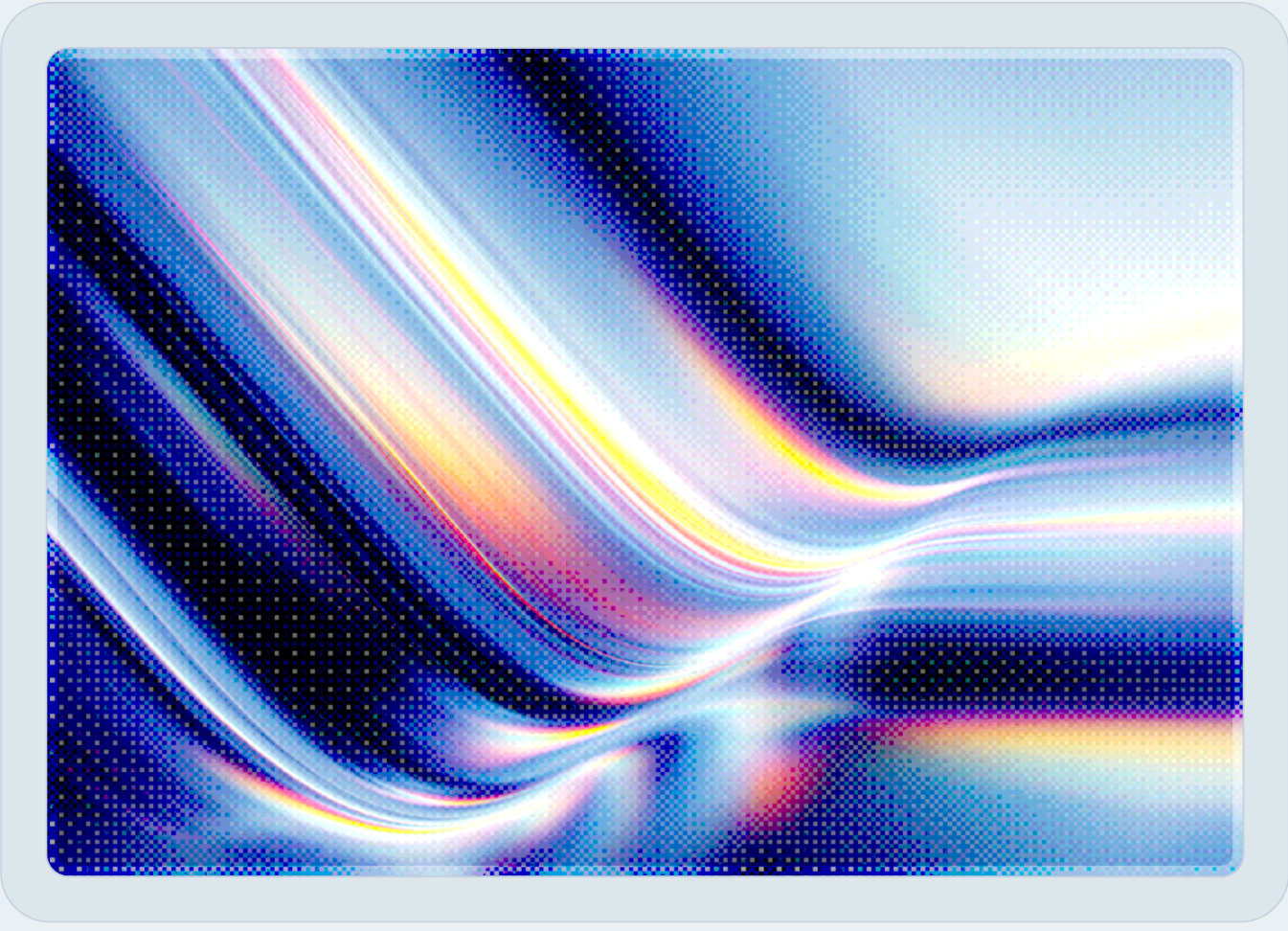
CSAI Foundation | Cloud Security Alliance

AI-Built Zero-Day: Attackers Weaponize LLMs Against 2FA

Google GTIG's May 2026 Discovery Signals a Structural Shift in How Adversaries Find and Exploit Authentication Flaws

2026-05-25

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On May 11, 2026, Google's Threat Intelligence Group (GTIG) publicly disclosed the first confirmed case of an AI-generated zero-day exploit deployed in the wild: a Python script that exploited a semantic logic flaw to bypass two-factor authentication in a widely deployed open-source web administration tool, as reported independently by multiple security publications covering the disclosure. [1][2][7]
- The vulnerability was not a memory-safety crash but a hardcoded trust assumption — precisely the category of subtle semantic error that traditional static analysis misses, because the code is functionally correct in isolation but semantically flawed within the broader authentication flow. [2]
- AI-specific artifacts in the code — educational docstrings, a hallucinated CVSS score, textbook-clean Pythonic structure — gave GTIG the analytic basis to assess with high confidence that the actor likely leveraged a large language model; no specific model was publicly identified. [1][7]
- This discovery does not stand alone: it reflects the leading edge of a broader AI-enabled authentication bypass ecosystem that includes industrial-scale phishing kits (Tycoon 2FA, BlackForce [10], EvilTokens, Kali365) and a post-disruption pivot to device code phishing responsible for more than seven million attacks in a single four-week period. [3][4][5][9]
- Conventional time-based one-time passwords assume attacks arrive at human speed; adversary-in-the-middle (AiTM) infrastructure — particularly implementations where AI automates session orchestration in place of a human operator — intercepts session cookies in real time and renders OTP-based 2FA structurally inadequate against a capable, motivated adversary. [6][10]
- Security teams should treat the GTIG disclosure as a forcing function: accelerate migration from TOTP-based 2FA to phishing-resistant FIDO2 credentials or passkeys, implement continuous session-level anomaly detection, and align authentication posture to zero-trust architecture principles that verify every session independently of MFA completion. [14]

Background

On May 11, 2026, Google's Threat Intelligence Group published a disclosure that many in the security community had anticipated as an eventual milestone: a confirmed case of a threat actor using an AI model to write a functional zero-day exploit and stage it for mass exploitation. [1] The target was a popular but unnamed open-source web administration tool. The flaw was not discovered through exhaustive fuzzing or memory-corruption scanning; it was a logic error in the application's authentication flow that may have passed prior code reviews precisely because it was syntactically valid — the code is, by any structural measure, correct, even as it creates a flawed trust boundary in the broader authentication context.

The exploit, a Python script designed for broad automated deployment, would have allowed any attacker already in possession of valid credentials to step past the 2FA challenge entirely. Google's analysts worked with the vendor to patch the vulnerability before the campaign launched, preventing the planned mass exploitation event. [2] But the significance of the disclosure extends well beyond this specific case. John Hultquist, chief analyst at GTIG, stated plainly: "There's a misconception that the race to AI vulnerabilities is imminent. The reality is it has already started. For every zero-day we can trace back to AI, there are probably many more out there." [1]

This note examines the technical and strategic dimensions of the GTIG incident, situates it within the current AI-enabled authentication attack landscape, and provides actionable guidance for security teams reassessing the reliability of their 2FA deployments in an era when the exploit development cycle has shortened dramatically.

Security Analysis

How an AI Finds What Scanners Miss

The particular danger of the vulnerability GTIG described is not its severity in isolation but its invisibility to conventional tooling. The flaw was a hardcoded trust assumption embedded in the application's authentication flow — a "if this object is present, trust it" shortcut that a developer introduced without apparent malicious intent. [2] Traditional static analysis tools and software composition analysis scanners evaluate code primarily for structural patterns: known-bad function calls, memory management errors, injection sinks, and dependency version matches. A line of code that enforces a trust condition incorrectly will typically pass all of those checks because the code is, by any syntactic measure, correct.

What distinguishes a capable language model from a scanner in this context is the ability to reason about semantic intent across a code path rather than pattern-match against individual lines. An LLM prompted to review authentication logic can ask, in effect, whether the trust model is internally consistent — whether what the code says it requires and what it actually enforces match. That reasoning capacity is precisely what surfaced this flaw where conventional review had not.

GTIG identified the exploit as AI-generated through a cluster of distinctive artifacts: organized educational docstrings that explained the code's purpose in tutorial-like detail, a hallucinated CVSS score presented as factual, and a textbook-clean Pythonic code structure that human exploit authors rarely produce. [1][7] GTIG assessed with high confidence that the actor likely leveraged an AI model — a probabilistic conclusion consistent with the artifact evidence rather than definitive proof of a specific system; contemporary reporting indicates analysts ruled out Gemini and considered commercial or jailbroken models as the leading working hypothesis. [7] These fingerprints are not proof of a specific model but are highly characteristic of large language model output, particularly from models prompted to produce well-commented, production-quality tool code.

The Surrounding Ecosystem: AI Accelerates the Full Attack Stack

The GTIG disclosure is best understood not as an isolated incident but as a milestone on a trajectory that the broader authentication threat landscape has been building toward for two years. AI has progressively automated and scaled each phase of the authentication bypass kill chain: initial credential collection, real-time OTP interception, session cookie theft, and now zero-day vulnerability exploitation to eliminate authentication requirements entirely.

The phishing-as-a-service market reached industrial scale in 2025. Tycoon 2FA, an adversary-in-the-middle phishing kit first observed in August 2023, had by mid-2025 become a dominant mechanism for bypassing MFA at volume. Microsoft's analysis found it responsible for approximately 62 percent of AiTM-category phishing campaigns the company tracked, with the platform linked to roughly 96,000 confirmed victims, with particularly severe impact across enterprise sectors. [3][8] AiTM infrastructure of this type does not defeat 2FA through brute force or credential guessing. It inserts an attacker-controlled reverse proxy between the victim and the legitimate service, relays the authentication challenge in real time, and captures the resulting session cookie after the user has successfully completed the MFA step. The victim experiences a normal login; the attacker receives a valid session token that remains usable independent of the MFA that generated it.

BlackForce, documented in December 2025, operates as a Man-in-the-Browser platform targeting MFA-protected accounts and is offered commercially at approximately €200–€300 per campaign. [10] Europol and Microsoft disrupted Tycoon 2FA in a coordinated operation in early 2026. [8] The disruption created a brief market vacuum that successor platforms filled within weeks. EvilTokens, a device code

phishing kit documented by Sekoia in March 2026, shifted the attack vector from credential interception to OAuth device code abuse — a technique that exploits Microsoft's authentication flow for input-constrained devices to grant account access without requiring the attacker to intercept credentials at all. [5] Barracuda's threat detection team observed more than seven million device code phishing attacks in a single four-week period ending April 23, 2026. [4] In May 2026, the FBI issued a Public Service Announcement warning organizations of Kali365, a Phishing-as-a-Service platform emerging in April 2026 and distributed via Telegram, which enables attackers to harvest Microsoft 365 access tokens and bypass MFA without ever touching the user's credentials directly. [9]

AI integration into these platforms spans several capability layers. Automated target profiling identifies high-value accounts within harvested credential sets and selects appropriate lure templates. Natural language generation produces phishing content that is often indistinguishable from legitimate corporate communications, substantially closing the quality gap that previously made AI-generated lures detectable. Real-time session orchestration routes authentication challenges faster than human operators could manage. The practical effect is that what once required a skilled attacker managing a live session now runs with minimal human involvement, enabling simultaneous attacks against hundreds or thousands of targets from a single operator. [6][10]

Why 2FA's Security Assumption Has Broken Down

The security proposition of time-based one-time passwords rests on a specific timing assumption: a valid OTP expires before an attacker can intercept and replay it. That assumption was reasonable when attacks required manual human-in-the-middle operation. AiTM infrastructure — and particularly implementations where AI replaces the human operator in real-time credential relaying, collapsing the interception window to sub-second timescales — operates at a speed that conventional OTP expiry windows cannot address. [6] The session cookie captured after a user completes an MFA challenge remains valid for hours or days depending on the application's session management configuration. Once captured, it completely circumvents the 2FA mechanism for all subsequent requests, because the application treats the authenticated session as established.

Against a zero-day exploit of the type GTIG described, the timing assumption does not apply at all. If an attacker can bypass the 2FA verification step in the authentication logic itself — bypassing the check rather than defeating the timing — then OTP validity is irrelevant. The exploit does not need to win a race against the one-time password window; it routes around the check that enforces the window entirely. This distinction matters because it defines the boundary between problems that improved OTP delivery or shorter token expiry can address and problems that require architectural change in the authentication mechanism itself.

Phishing-resistant authentication standards — FIDO2 hardware keys and passkeys based on the WebAuthn specification — defeat both attack classes when properly implemented. The cryptographic binding of a FIDO2 assertion to a specific origin domain prevents replay against a spoofed domain regardless of how quickly an attacker intercepts the exchange. And a logic-flaw bypass of the type GTIG identified would, in a FIDO2-enforced flow, still require defeating the origin-bound cryptographic challenge rather than simply satisfying a trust assumption in application code. Neither AiTM infrastructure nor the class of semantic logic exploits AI is beginning to surface can trivially bypass a properly implemented FIDO2 deployment.

The Zero-Day Discovery Acceleration Problem

The GTIG incident connects directly to a broader pattern documented in recent months: AI is reducing the cost and expertise required to find exploitable vulnerabilities in production software. DARPA's AI Cyber Challenge, concluded at DEF CON 33 in August 2025, demonstrated that AI-assisted systems developed by competition teams could reportedly process tens of millions of lines of code, validate dozens of synthetic vulnerabilities, and uncover real-world flaws at an average task cost measured in the low hundreds of dollars — a fraction of the traditional cost of manual vulnerability research. [11] As the cost of vulnerability discovery approaches the cost of a single cloud computing session, the historical assumption that zero-day development requires rare expertise and significant investment no longer holds.

For authentication-specific logic flaws in particular, this acceleration is especially consequential. Authentication code is widely shared: popular open-source web frameworks, identity libraries, and administration tools are deployed across tens of thousands of organizations. A single logic vulnerability in a widely deployed component, once discovered and weaponized, translates to a reusable attack against any organization running that component without the attacker needing to conduct organization-specific reconnaissance. The GTIG case — a single exploit staged for mass exploitation against an unnamed but widely deployed tool — illustrates exactly this leverage.

Recommendations

Immediate Actions

Organizations currently relying on TOTP-based 2FA as their primary authentication barrier should treat this incident as a trigger for security architecture review rather than a footnote to an existing program. The first immediate step is to identify where TOTP-based 2FA serves as the sole second factor for high-

value access: administrative consoles, cloud provider accounts, identity provider administration, code repositories, and financial systems. Those environments carry the highest consequence if a session bypass or logic-flaw exploit reaches them, and they represent the priority queue for migration to phishing-resistant credentials.

Incident response teams should simultaneously review session management configurations across critical applications, ensuring that session token lifetimes are minimized and that abnormal session behavior — concurrent sessions from geographically inconsistent locations, session use following credential change, or access patterns inconsistent with established baselines — triggers review or revocation rather than silent continuation.

Short-Term Mitigations

For systems where FIDO2 migration cannot be completed immediately, layering controls reduces the exploitation surface without requiring credential infrastructure changes. Conditional access policies that restrict authenticated sessions to known device states, compliant devices, or specific network ranges limit the utility of a captured session cookie to an attacker operating from infrastructure outside those bounds. Token-binding mechanisms, where supported, cryptographically tie access tokens to the client connection, preventing replay from a different session context.

Detection engineering should incorporate the behavioral signatures of AiTM and device code phishing at the identity layer. Microsoft and independent researchers have published detectable indicators for Tycoon 2FA, EvilTokens, and Kali365, including characteristic OAuth request patterns, anomalous device registration events, and token issuance from unexpected client applications. [3][5] Security teams that have not mapped these indicators to SIEM detection logic should prioritize that work given the documented volume of current attacks.

Strategic Considerations

The GTIG disclosure should inform a longer-term shift in how organizations assess the residual risk of 2FA deployments. The security community has treated TOTP-based 2FA as a substantial barrier for more than a decade, and it remains valuable against many attack classes. But AI-enabled attackers are systematically eroding the assumptions that make it sufficient: the speed of interception, the cost of zero-day discovery, and the scalability of automated session exploitation have all moved in directions that favor the attacker.

FIDO2 passkeys, currently deployable for most modern cloud-based enterprise applications via identity providers including Microsoft Entra, Okta, and Google Workspace, represent the most direct architectural response to phishing-based 2FA bypass, particularly for organizations most exposed to the

AiTM threat class. Migrating administrative and high-privilege access first, followed by broad workforce rollout, aligns effort with consequence. A parallel investment in zero-trust network architecture — continuous session verification, microsegmentation, and application-layer access policies that enforce least privilege at each request rather than at login — addresses the session persistence problem that AiTM exploits even when 2FA completion succeeds.

Security teams should also monitor for the class of semantic logic vulnerabilities the GTIG case exemplifies. Static analysis tools are beginning to incorporate LLM-assisted reasoning that can evaluate authentication flow consistency rather than just structural patterns. Integrating these tools into CI/CD pipelines and applying them specifically to authentication and session management code represents a defensive use of the same capability that attackers have begun to weaponize.

CSA Resource Alignment

The threat pattern described in this note maps directly to several CSA frameworks and programs. CSA's MAESTRO framework for agentic AI threat modeling addresses the conditions under which AI systems — including adversarially deployed ones — interact with authentication and identity infrastructure. MAESTRO's Layer 1 (Foundation Models) and Layer 3 (Agent Frameworks) threat categories are directly implicated when a large language model generates functional exploit code that interacts with authentication services, and organizations adapting MAESTRO for defensive purposes can use its threat taxonomy to enumerate authentication-layer risks from AI-enabled adversaries. [12]

The AI Controls Matrix (AICM) v1.0 provides control mappings specifically relevant to identity and access management in AI-enabled environments. The AICM's identity domain controls — covering authentication requirements for AI-adjacent systems, session governance, and privileged access management — offer a structured baseline for organizations assessing whether their current 2FA posture meets the threat level documented here. CSA's Zero Trust guidance provides the architectural foundation for the session-level verification and least-privilege access enforcement that this note recommends as strategic mitigations. Organizations should also consult the CSA STAR for AI program criteria, which include authentication and access control requirements that reflect the current threat environment, when assessing the security posture of AI-adjacent administrative interfaces. [13]

References

[1] The Hacker News. "Hackers Used AI to Develop First Known Zero-Day 2FA Bypass for Mass Exploitation." The Hacker News, May 2026.

[2] SecurityWeek. "Google Detects First AI-Generated Zero-Day Exploit." SecurityWeek, May 2026.

[3] Microsoft Security Blog. "Inside Tycoon2FA: How a Leading AiTM Phishing Kit Operated at Scale." Microsoft, March 4, 2026.

[4] Barracuda Networks. "Threat Spotlight: Device Code Phishing Is on the Rise with 7 Million Attacks in Four Weeks." Barracuda Networks Blog, April 23, 2026.

[5] Sekoia Threat Detection & Research. "New Widespread EvilTokens Kit: Device Code Phishing-as-a-Service – Part 1." Sekoia Blog, March 30, 2026.

[6] WorkOS. "How Attackers Are Bypassing MFA Using AI in 2026." WorkOS, 2026.

[7] BleepingComputer. "Google: Hackers Used AI to Develop Zero-Day Exploit for Web Admin Tool." BleepingComputer, May 11, 2026.

[8] CybersecurityNews. "Tycoon 2FA Phishing Kit Disrupted by Microsoft, Europol and Partners." CybersecurityNews, 2026.

[9] Internet Crime Complaint Center. "Kali365 Phishing-as-a-Service Kit Hijacks Microsoft 365 Access Tokens." FBI IC3 Public Service Announcement, May 21, 2026.

[10] The Hacker News. "New Advanced Phishing Kits Use AI and MFA Bypass Tactics to Steal Credentials at Scale." The Hacker News, December 12, 2025.

[11] DARPA. "AI Cyber Challenge (AIxCC)." DARPA, 2025.

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

[13] Cloud Security Alliance. "STAR for AI Program." Cloud Security Alliance, 2026.

[14] Cybersecurity and Infrastructure Security Agency. "Implementing Phishing-Resistant MFA." CISA, 2022.