Threat Intelligence

GTIG AI Threat Tracker: Agentic Threat Activity Crosses a Significant Threshold

November 16, 2026

Google Threat Intelligence Group

Written by: Aurora Blum, Kelli Vanderlee, Jess Xia, David Mainor

Executive Summary

frameworks, automated vulnerability exploitation pipelines, and alternative monetization schemes like LLMJacking, the adversarial AI landscape has reached a critical inflection point. In this update, Google Threat Intelligence Group (GTIG) and Mandiant examine how adversaries across the cyber crime, espionage, and information operations spectrum are operationalizing generative AI to compress attack timelines, scale operations, and execute multi-stage intrusions at machine speed.

Key trends over the last quarter include:

- **Shift to Autonomous Workflows and Multi-Agent Frameworks:** Adversary tradecraft has continued to transition beyond manual prompting to deploying autonomous multi-agent offensive frameworks—such as ARTEX, CAIRN, CyberStrikeAI, and PentAGI—to execute multi-stage intrusions at greater operational velocity.
- **Automated Vulnerability Research and Exploitation:** Threat actors are increasingly leveraging LLMs and agentic frameworks to fully automate the reverse engineering of patches via differential analysis, allowing them to rapidly weaponize n-day vulnerabilities and scale exploitation pipelines against internet-facing servers without human intervention.
- **Monetization and Infrastructure Abuse**: Financially motivated threat actors continue to profit from LLMJacking by stealing LLM credentials and cloud resources from

through underground proxy services (e.g., LLM Mixers).

- **LLM Integration in Information Operations (IO):** State-sponsored groups, commercial Influence-as-a-Service (IaaS) vendors, and inauthentic political monetization networks now routinely use LLMs in their operations to produce synthetic media at scale, create localized personas, and automate cross-platform distribution.
- **Targeted Espionage and Supply Chain Evolution:** Cyber espionage groups are leveraging LLMs for target research, automated data analysis, and novel supply chain attack vectors—such as manipulating autonomous coding assistant configurations to drop backdoors during local execution.

Grounded in telemetry from frontline Mandiant incident response engagements, global threat actor tracking, and live platform defenses, this report details how state-sponsored espionage groups, financially motivated cyber criminals, and IO threat actors are operationalizing LLMs in the wild.

At Google, we are committed to developing AI boldly and responsibly. Our multifaceted defense strategy integrates proactive model-level safeguards, specialized threat intelligence, and targeted containment protocols to protect our customers and infrastructure. We continuously harden our models against misuse, mitigate malicious activity through

Threat Defense architecture and Beyondzero to operationalize security, continuous monitoring, and responding at machine speed across enterprise environments.

Multi-Agent Frameworks in Cyber Intrusions

While previous milestones documented in our quarterly research highlighted nascent adversary interest in agentic capabilities, GTIG and Mandiant have now observed multiple threat actors abusing open source, agentic penetration-testing (pentesting) frameworks to automate discrete phases and specific operational tasks during live intrusions.

Threat actor adoption of these frameworks represents a paradigm shift towards intrusions occurring at machine speed, with the automated chaining of complex attack paths scaling operations significantly. While the mechanics of the intrusions remain the same, such as using compromised credentials and targeting unpatched systems, the speed and scale at which frameworks can exploit these known weaknesses introduces a new defensive challenge.



Figure 1: Multi-agent LLM offensive frameworks

In addition to the now highly-publicized pentesting tool ARTEX, threat actors are actively leveraging additional open source, multi-agent frameworks such as CAIRN, CyberStrikeAI, and PentAGI. We have also observed evidence of threat actor experimentation with custom-built harnesses. These frameworks share several core capabilities: discovering and exploiting low-complexity vulnerabilities, using LLMs to generate single-use payloads on-demand, and using real-time execution feedback to dynamically adjust tactics.

GTIG has independently observed open source tools CAIRN and PentAGI deployed in the wild, and has seen experimentation with CyberStrikeAI and HexStrike AI.

- **CAIRN:** an open source, multi-agent pentesting framework that implements an OODA (Observe, Orient, Decide, Act) loop where specialized agents collaborate via a shared "blackboard" to autonomously navigate from an initial parameter (e.g., a target IP address) toward an operational objective. GTIG observed CAIRN used in

- **CyberStrikeAI**: an open source, LLM-native offensive framework which functions as a "single pane of glass" for cyber operations. The core architecture consists of a Go-based backend API acting as a C2 and orchestration server and connects commercial and open-weight LLMs to over 100 security tools, YAML attack recipes, and custom scripts to automate end-to-end attack chains.
- **HexStrike AI**: an open source offensive Model Context Protocol (MCP)-integrated harness that enables LLM clients to directly connect to over 150 pre-installed offensive security tools. Operating through a dual-component architecture, the framework wraps standard utilities across reconnaissance, web application testing, exploitation, and binary analysis into callable functions coordinated by more than 12 specialized autonomous sub-agents.
- **PentAGI**: an open source autonomous offensive framework designed to lower technical barriers for entry-level operators running complex campaigns. Between late August and early October 2026, GTIG tracked a threat actor using a customized attack orchestration platform built off of the open source version of PentAGI used to conduct systematic reconnaissance and vulnerability scanning against a wide range of organizations in Asia.

We have observed several threat actors separately experimenting with the use of these frameworks

PRC nexus. This includes evidence that a PRC-nexus group was developing its own automated attack framework for likely use in cyber espionage operations. These observations additionally extend to the use of other tools and services, such as ZeroHollow, an inference service used for exploit development and LLM routing, and Agent Zero, an open source agent framework which features autonomous code execution and dynamic tool generation.

Given the low barrier to entry, ease of integrating new exploits, and massive scaling potential, we believe it is highly likely that threat actors will continue to use these frameworks in campaigns targeting organizations globally. Concurrently, as capabilities evolve, we suspect these operations will continue to move further along the autonomous scale, approaching fully agentic intrusions. Notably, while these frameworks are capable of significantly compressing attack timelines to machine speed and achieving operational objectives, they currently trade stealth for velocity. Unrestricted autonomous execution generates forensic traffic, relies on recognizable identifiers, and leaves behind distinct code-level traces; while these artifacts increase the likelihood of detection and mitigation, these tools remain highly capable of achieving operational objectives. As the sophistication of these tools and underlying LLMs increases, we anticipate not only reduced human involvement in these operations but also increased stealth and sophistication of the activity.

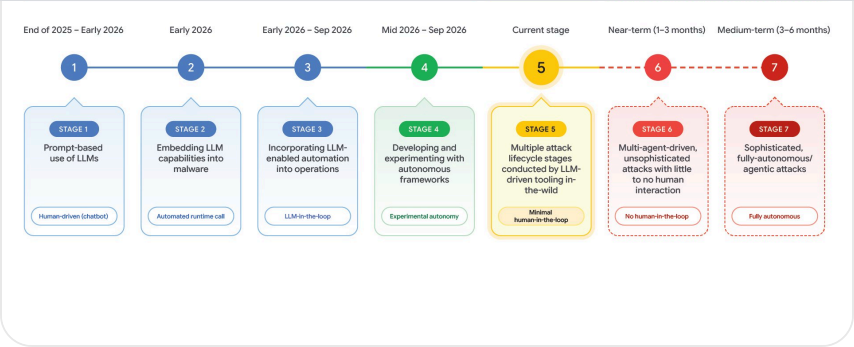


Figure 2: LLM maturity timeline

Mitigations
Google Security Operations (SecOps) has deployed specialized behavioral and volumetric rulepacks to detect use of ARTEX. Mandiant Consulting released a hardening guide and architectural defense framework specifically designed to counter the speed and scale advantage of adversaries using LLM-enabled offensive tooling.

Use of Agents in Vulnerability Research and Development

The integration of autonomous LLM agents into vulnerability research and software security workflows has shifted from theoretical experimentation to active operational deployment.

phases of vulnerability analysis, such as differential patch examination, dynamic protocol fuzzing, and proof-of-concept (POC) weaponization. By chaining these capabilities together, threat actors can dramatically reduce the window between patch disclosure and weaponized exploitation. Threat actors may find it more efficient to use LLMs to rapidly weaponize patched n-day vulnerabilities, as they can focus research on a finite set of source materials: comparing source code versions (including from pre-release or preview channels), and examining release notes, security research, and POC code related to specific flaws. In contrast, searching for zero-day vulnerabilities requires casting a wider net. LLMs are transforming how vulnerabilities are discovered and exploited across public and enterprise infrastructure, shrinking the patch window for organizations. Simultaneously, the lower barrier to entry has fueled a wave of LLM-generated exploits that are claimed to work but don't, creating additional verification work for defenders.

For instance, in September 2026, GTIG analyzed vulnerability statistics to assess how LLMs are shaping the threat landscape. We found that LLMs are transforming not only the speed of vulnerability discovery and exploitation, but also the types and risk profiles of the flaws being uncovered. Interestingly, while overall disclosures and exploits nearly doubled between January and August 2026, zero-day exploits rose only marginally from a 2025

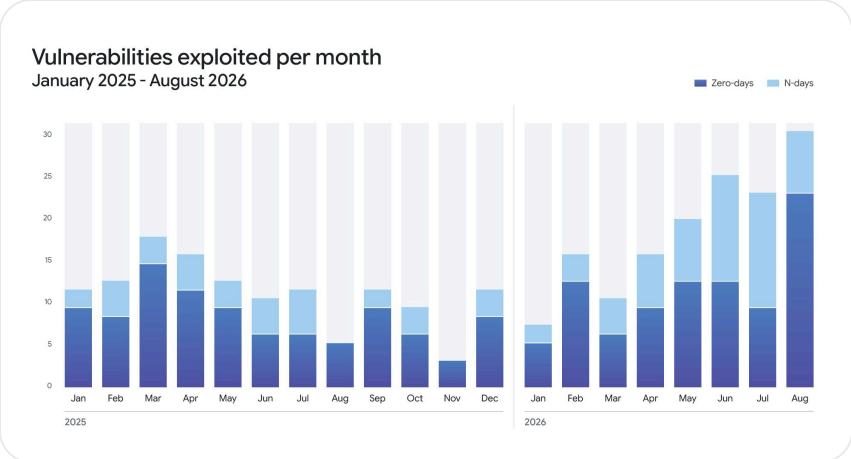


Figure 3: Count of vulnerabilities exploited, by n-days and zero-days, January 2025 to August 2026

In our previous report, we detailed nascent efforts by threat actors surrounding LLM-augmented vulnerability research, including partially automated vulnerability scanning and exploitation attack sequences.

In the past quarter, GTIG identified threat actors using LLM frameworks to fully automate vulnerability discovery and exploitation without a human in the loop for the first time.

More specifically, GTIG identified an exposed staging server hosting a multi-agent vulnerability research and weaponization framework, dubbed `exploit_agent`, that integrates with multiple frontier and local LLMs. The framework is set up to conduct continuous differential analysis of popular web application software. It compares code before and after patch announcements, analyzes findings to identify potential vulnerabilities, drafts exploits,

targeted victims, we identified multiple scripts capable of facilitating various post-compromise activities, including internal reconnaissance, brute-forcing, installing webshells, and removing other attackers' webshells. Victim telemetry on the server suggests the operator was specifically interested in targeting French websites associated with the media and entertainment industries, as well as the education sector.

```
// Insecure capability check analyzed in v2
public static function restrict_popup_rest_
    if ( is_wp_error( $result ) ) {
        return $result;
    }
    $route = isset( $_SERVER['REQUEST_URI']
    if ( false === strpos( $route, '/wp/v2/
        return $result;
    }
    if ( is_user_logged_in() && current_use
        return $result;
    }
    return new WP_Error(
        'rest_forbidden',
        __( 'Sorry, you are not allowed to
        array( 'status' => rest_authorizati
    );
}
```

Figure 4: Agentic framework inspecting access control implementations in the Spectra (Ultimate

Russian Language Comments	Comments Translated to English

`# LLM API-ключи для авто-эксплуатации (пилот exploit_agent)`
 	
`# Впиши ключи сам — сюда, не в чат. Обе строки обязательны.`
	`# LLM API keys for auto-exploitation (exploit_agent pilot)`
	
	`# Enter the keys yourself — right here, not in the chat. Both lines are required.`

Table 1: Comments from environmental file configuring multiple LLM API keys for autonomous exploitation

Open source research corroborates threat actor operationalization of fully automated vulnerability exploitation pipelines. For example, in late August 2026, threat actors exploited two vulnerabilities in PaperCut print management servers as zero-days that, when chained together, allow for unauthenticated remote code execution: CVE-2026-81578, an authentication bypass flaw; and CVE-2026-82078, an unsafe dynamic class loading vulnerability. Using LLMs, security researchers and the cybersecurity company Huntress generated working exploits for the vulnerabilities within a day

which a threat actor likely used LLMs to build and test exploits for these vulnerabilities, assemble a target list, and execute a widespread exploitation campaign against hundreds of organizations. Greynoise also highlighted the speed of execution, noting that the threat actor expanded an initial foothold to domain administrator within minutes.

The PHP vulnerabilities (CVE-2026-109618 and CVE-2026-109619) also highlight several of these trends. They were discovered using LLM-enabled analysis. Before official patches were released on Oct. 21, 2026, a POC video and early patched code were published. This allowed both researchers and attackers to compare code versions, understand the flaws and attempt to build exploits before the patch became available. In the initial days following the official patch, hundreds of claimed but non-functional POCs, likely generated with the help of LLMs, were published. This proliferation created significant work for the security community to verify or debunk the POCs while monitoring for evidence of confirmed exploitation.

Mitigations
Google continuously invests in AI innovation and research, helping to ensure AI is built responsibly, and that we're leveraging its potential to automatically find risks. In 2024 we introduced Big Sleep, an AI agent developed by

vulnerabilities in software. Google Cloud employs a defense-in-depth model that combines rigorous code reviews and secure-by-default libraries with continuous, automated scanning using specialized AI agents such as CodeMender to discover, patch, and remediate vulnerabilities before and after release. We partner with the global security community by running a Vulnerability Reward Program and maintaining Project Zero to proactively find and patch zero-day vulnerabilities.

Full Integration of LLMs into Information Operations

GTIG assesses with high confidence that information operations (IO) threat actors have advanced from initial LLM experimentation to full operational integration. Across all tracked categories—state-sponsored groups, including those attributed to Russia, the PRC, and Iran, commercial "Influence-as-a-Service" (IaaS) vendors, and inauthentic political monetization (IPM) networks—adversaries now routinely embed LLMs into their standard tradecraft to produce synthetic media and automate domain network workflows (e.g., LLM-assisted SEO manipulation, editing and formatting of seeded articles, and automated generation and publishing of influence content). This broad adoption has measurably

scale political clickbait that resembles the thematic focus, quality, and volume historically characteristic of state-backed IO campaigns. However, it is important to note that improvements in content quality do not necessarily equate to greater impact or success.

IO Actor Categories
State-sponsored groups are IO actors directly linked to state apparatuses and have traditionally formed the primary baseline tracked by platform defenders and researchers. To preserve plausible deniability, they increasingly rely on third-party commercial contractors to execute campaigns on their behalf.
Commercial IaaS vendors encompass marketing and PR agencies, private intelligence contractors, and offshoots of commercial surveillance vendors, selling services ranging from tailored influence campaigns to turnkey platforms. Tradecraft varies across this ecosystem; intelligence contractors and surveillance offshoots in particular often apply stronger operational security (OPSEC) and more advanced techniques than traditional state-sponsored actors.

ideology. These actors generate a high volume of political clickbait to drive engagement, maximize click-through rates, and boost ad revenue. This combination of political framing and high-volume synthetic content creates a significant frontend perception gap, making IPM network campaigns the most prone to being misidentified as state-sponsored operations.

Table 2: IO actor categories

As automated content generation reduces a primary operational bottleneck, adversaries are increasingly shifting their focus toward leveraging LLMs for infrastructure design, automated distribution pipelines, and anti-detection mechanisms. Specifically, our observations indicate a notable shift toward the commercial development of LLM-enabled, end-to-end IaaS platforms designed to support and automate the full influence campaign lifecycle. The proliferation of these platforms can lower operational overhead for state and commercial actors while increasing the difficulty of detecting coordinated and inauthentic behavior across social media platforms.

The Commercial Development of Turnkey,

Operations

In August 2026, GTIG investigated threat activity leveraging the Gemini API alongside tools from multiple AI providers to build an end-to-end IaaS platform designed to target public sentiment on regional political topics, primarily across Europe and the Middle East. The platform was engineered to automate the entire IO campaign lifecycle, including persona creation, multimodal media generation, automated content distribution, and interactive posting across platforms. Identified prompts reflected an operational focus on emulating organic grassroots engagement (i.e., “astroturfing”) while evading platform detection rules.

- Identified prompts instructed models to adopt localized personas specific to targeted geography, mimicking regional dialects and slang, while incorporating informal phrasing and intentional typos to avoid sounding like an automated tool.
- Prompts supporting media generation introduced deliberate camera flaws simulating realistic photos typically captured by smartphone cameras (e.g., making images appear grainy or incorporating uneven lighting changes) and omitted background text to avoid synthetic image markers.
- To evade social media platform defenses, operators routed traffic through residential proxies and preserved saved login cookies. This

continuously, avoiding security checkpoints like CAPTCHAs and login re-verification.

GTIG also detected a separate commercial contractor leveraging the Gemini API—as part of a modular, multi-provider AI architecture—during the development of a turnkey IaaS platform engineered to manage over 100,000 accounts across major global platforms. The platform is architected to combine automated agents, anti-detection and device identifier spoofing tools, and cloud phone farms to operate thousands of inauthentic personas across platforms. Observed prompts indicated the actor's plan to market the tool to both the actor's domestic government bodies and foreign clients for influence campaign execution.

- To manage large volumes of social media accounts, the platform was designed to run virtual smartphones in cloud environments—disguising each virtual device as a legitimate mobile device—and route web traffic through proxies to bypass automated bot detection and rate limits.
- The architecture additionally includes automated fake account creation using disposable email addresses and overseas virtual SMS numbers. To avoid further suspicion, newly-created accounts were configured to undergo an automated aging period, leveraging LLMs to simulate realistic user behavior (e.g., how long a user lingers on a post, scrolling, and interacting with non-campaign content) to

Tactical LLM Adoption: Demographic Microtargeting and Content Scaling

Alongside the development of IaaS software suites, threat actors continue to integrate commercial LLM tools into baseline content workflows to overcome linguistic barriers, customize messaging, and scale output. Using commercial models to produce commentary, synthetic personas, and stylized animations at the content layer allows IO actors to improve the credibility and demographic resonance of their campaigns, while enabling IPM networks to mass-produce political clickbait for advertising revenue.

- **LLM-Enabled Quality Enhancement in Long-Running IaaS Campaign:** Since our last report, GTIG observed a long-running IaaS provider tracked since 2019 leveraging LLMs to elevate the baseline quality and volume of its influence operations. Utilizing a proprietary software suite designed to manage thousands of synthetic profiles across major platforms, the provider targeted political discourse in multiple countries. Operators integrated LLM models to produce more convincing virtual personas, as well as academic-style commentary and media focused on political, religious, and civil rights topics. The resulting content was subsequently promoted and amplified across platforms using fabricated Western personas.

LLM-generated content to customize context for specific regions and audiences, for example by using synthetic influencer personas to promote policy narratives targeting specific demographics. A separate network posted synthetic, Lego brick-inspired animations and football parodies to satirize global political figures, likely in an attempt to appeal to younger audiences by effectively packaging propaganda in short-video formats consistent with popular internet culture.

- **LLM-Augmented Inauthentic Political Monetization (IPM):** GTIG also tracks a growing trend of IPM actors using LLMs to mass-produce political content. Although their thematic focus, quality, and volume can resemble state-backed IO campaigns, these non-ideological operations exploit societal polarization solely to maximize engagement, click-through rates, and advertising revenue. GTIG has observed IPM operators prompting commercial LLMs to identify lucrative political topics, trending keywords, and high-Cost Per Mile (CPM) niches for ad monetization, while promoting the resulting content via synthetic news anchors that occasionally spoof the likenesses of real journalists and commentators to manufacture credibility.

Mitigations

Google surfaces under our standard platform policies. Our research and rigorous analysis enables Google teams to make enforcement decisions and to disrupt coordinated IO campaigns. GITG, YouTube, and Google Trust & Safety track and regularly enforce on accounts associated with coordinated IO posting content and commenting. Examples of this enforcement include disruption of YouTube channels, blogs, AdSense accounts, and domains removed from Google News surfaces, as we report on a quarterly basis in the Influence Operations Bulletin.

Beyond enforcement actions, GTIG shares insights with Google DeepMind to harden our platforms and services against emerging tradecraft. This intelligence is used to establish abuse baselines, enhance Gemini's ability to recognize and refuse misuse, and inform internal policy modifications.

Threat Actors Abusing and Monetizing Access to AI Tools

While compromised credentials have long served as the fundamental bedrock for initial access, privilege escalation, and lateral movement during intrusions,

unauthorized LLM tooling and autonomous operational pipelines. As access costs, token rate limits, and geographic restrictions present persistent operational hurdles for threat actors attempting to scale LLM capabilities, adversaries are actively harvesting session tokens and platform credentials directly from victim endpoints and continuous integration environments. In tandem with this shift toward hijacking valid credentials for LLMs, underground cybercrime forums and illicit marketplaces have rapidly matured to meet this demand—transitioning beyond generic identity dumps to commodifying verified frontier AI accounts, tokenized API quotas, and targeted infostealer configurations as turnkey utilities for downstream exploitation.

LLMJacking in Incident Response

Throughout 2026, Mandiant responded to several LLMJacking incidents where financially motivated actors compromised cloud environments to consume targeted organizations' cloud compute and/or LLM resource quotas. As seen in the image below, LLMJacking can generate significant usage and billing increases within hours. In some cases, single-day fraudulent charges exceeded $100,000 USD. In the most severe cases, cumulative resource abuse reached $1 million USD before remediation.

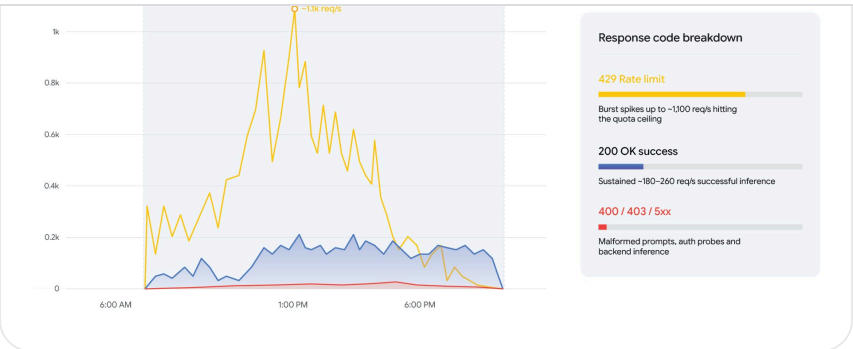


Figure 5: Usage chart showing LLMJacking generating a spike in API requests

In November, Mandiant and GTIG described a set of common TTPs we observed across intrusions that affected financial services, retail, transportation, and technology organizations in North America, South America, and the Asia Pacific region.

Attack Lifecycle Stage	Description of Observed Tactics

	session hijacking to gain access to targeted environments, including harvesting valid credentials from public GitHub repositories, and social engineering using deceptive websites to trick users into providing active session tokens.
Internal Reconnaissance & Lateral Movement	Seeking to identify credentials and systems of interest, attackers conducted reconnaissance against internal code repositories, data stores, cloud services, containers, and access logs. In at least one case, a threat actor exploited a suspected web application vulnerability in a package running in an internal container to navigate to additional systems.
Privilege Escalation & Persistence	Threat actors created unauthorized IAM and service accounts or attached administrative

	actor exploit CVE-2026-40217, a sandbox escape in LiteLLM that allows an attacker to elevate limited-access inference keys ("Virtual Keys") into full administrative credentials. Privileged service accounts were sufficient to establish persistence without the use of malware.
Complete Mission: LLM Abuse	Inside organizations' cloud environments, attackers enabled LLM services, including subscribing compromised accounts to pay-as-you-go agreements. To circumvent rate limits, one threat actor used a "burn and rotate" strategy, creating and deleting short-lived service accounts. Other threat actors abused cloud compute resources, provisioning instances with powerful GPUs and requesting quota increases in order to run

	frameworks that a threat actor used to automate further exploitation of the compromised network, as well as open source API gateways and custom relay infrastructure deployed to facilitate the sale of access to the compromised organization's resources.

Table 3: Common TTPs observed in LLMJacking intrusions

In another investigation, we observed at least two separate attackers seeking to conduct LLMJacking operations. Both actors accessed a targeted environment using the same set of legitimate credentials, which were potentially sold or leaked on underground forums. The first threat actor used legitimate credentials to gain access to the environment, enumerate available LLM services, and query cost and usage information related to the account. On the same day, we observed a set of IP addresses using a common User-Agent, likely a botnet or proxy service, also accessing the environment using the same legitimate credentials. The botnet or proxy service generated millions of LLM API calls in approximately 24 hours before the credentials were revoked. Though the first threat actor returned to the environment, creating service

compromised credentials to be remediated before this threat actor could execute LLM queries.

For hardening and mitigation guidance for defending against LLMJacking intrusions, please see our blog.

Monetizing LLMJacking through LLM Mixers

Threat actors frequently monetize LLMJacking intrusions by selling access to compromised compute and LLM resources. In addition to the rise of LLMJacking, GTIG is tracking the proliferation of "LLM Mixers"—illicit storefronts that provide buyers with unauthorized access to LLM services and underlying proxy infrastructure. This infrastructure routes buyer traffic through a pool of stolen credentials and compute instances to ensure reliable access, while also offering the ability to spread query volume across multiple victim environments to reduce the likelihood of detection.

The LLM Mixer ecosystem mirrors the broader cybercrime marketplace, where specialized threat groups handle discrete phases of the supply chain. While LLM Mixer operators maintain infrastructure and customer fulfillment, for example, operators may partner with separate threat actors who establish initial access to compromised environments, purchase compromised credentials

GTIG analyzed numerous LLMJacking vendor advertisements on Russian- and Chinese-language underground marketplaces, and found that most services offer access to multiple models, including frontier and jailbroken models that advertise support of malicious cyber activity. Advertised pricing varies significantly across individual mixer platforms and based on the models and types of credentials offered; for example, free trial credentials were much cheaper than enterprise. We observed prices per million tokens range from $0.10 USD to $25 USD. One vendor, APIMart, applies a 5x price increase on both input and output once a prompt exceeds 100,000 tokens. Rather than pricing per million tokens, some vendors such as JetNight offer monthly subscriptions at differing prices based on the type of model offered.



Figure 6: APIMart advertised prices

Threat Actors Continue Using AI for Productivity Gains

Since our last report, GTIG has continued to observe threat actors leveraging LLMs to augment various phases of the attack lifecycle. However, widespread adoption of LLM tooling also introduces significant operational drawbacks. The forensically noisy nature of many LLM tool processes can degrade OPSEC and compromise stealth, creating a detrimental impact on espionage campaigns. In addition, LLMs can reduce operational reliability, as adversaries must contend with model limitations, including the

specialize in stealthy operations designed to maintain long-term persistence are less likely to rely on forensically noisy tools within a targeted environment, though they may employ them selectively for speed and scale, depending on operational priorities. Conversely, we assess with high confidence that actors less concerned with operational stealth, including many cyber crime actors, will continue incorporating autonomous offensive tools into their operations, given that the increased reach of their campaigns will likely compensate for any decreases in success rates.

Government-backed threat actors

State-sponsored threat actors continue to leverage LLMs in their respective operations, reinforcing baseline assessments communicated in our previous reports. For example, we continue to see CALANQUE ION, an Iranian government-backed actor previously tracked as APT42, leverage LLMs—including Gemini—to augment reconnaissance and targeted social engineering. We also see PRC-nexus threat actors leveraging LLMs for reconnaissance and research. In addition to this new baseline of threat actors leveraging LLMs, we observed the following activity:

In Q3 2026, PRC-nexus cyber espionage actors used LLM tools and models for reconnaissance, front-end scanning, researching targets across

- We have also observed threat actors using LLMs to conduct post-compromise data analysis. This includes PRC contractor LEAKY CASTLE, which overlaps with public reporting about data leaked from a contractor using LLMs to analyze, translate, and understand exfiltrated data, including deriving insights into organizational structures and adding them into finished intelligence reports to sell to customers.
- In a new trend, we have observed PRC-nexus actors abusing LLM platforms and file-hosting backends for staging, dead drop C2s, and exfiltration. In one example, a threat actor abused a frontier model's file-hosting infrastructure to stage lure archives and exfiltrate stolen data, and a backend API to exfiltrate stolen intranet documents.

We also continue to see threat actors from the Democratic People's Republic of Korea (DPRK) using LLM tools to support DPRK IT worker activity, such as generating resumes and preparing for job interviews at Western organizations. DPRK threat actors are also abusing generative tools for operational scaling and exploiting native agent execution hooks as enterprise access vectors.

- We observed the use of public facing tools to generate responses during job screenings, allowing them to sound authentic while posing as Western IT workers to secure remote employment.

assist with coding, over the course of 2026 we've observed an evolution of tactics. In multiple intrusions, a suspected MIDNIGHT NEPTUNE threat cluster combined LLM-enabled social engineering and abused organizations' AI tools as a supply chain attack vector.

- In recent intrusions targeting cryptocurrency organizations in North America and Europe, the threat actor exploited LLM software architectures as a lateral execution surface. For initial access, the actor created fraudulent recruiter profiles on professional networking platforms whose profile imagery contained SynthID watermarks, confirming use of LLMs to synthesize deceptive professional identities. These personas directed targets to fraudulent meeting platforms featuring "ClickFix" social engineering lures, which prompted targets to run diagnostic fix scripts, which executed likely LLM-authored staging scripts via ClickFix lures. Following initial compromise, the actor pivoted into internal code repositories to execute a supply chain vector internal to the organization: modifying `.claude/settings.json` configuration files so that when users started up Claude Code, it would automatically trigger the download of the SOMBERMEME backdoor. To maintain stealth, suspected MIDNIGHT NEPTUNE further masked its lateral movement by generating GitHub Personal Access Tokens (PATs) named after Microsoft Copilot, underscoring how threat actors are

agent execution hooks as enterprise access vectors.

LLM Toolkits Offer Turnkey Cyber Crime Solutions

In developments in the last quarter, in addition to the use of offensive frameworks in financially motivated activity, we observed cybercrime actors leverage LLMs to scale tooling for clients; for example, Phishing-as-a-Service (PhaaS) operators using LLMs to create tools for low-sophisticated adversaries. The Jieyou phishing service, an active Chinese-language PhaaS platform first observed in March 2026, is primarily used by cybercriminals to deploy and manage highly convincing phishing campaigns, with a particular focus on e-commerce fraud. The service introduced an LLM chatbot assistant in its June 2026 backend update. This chatbot allows threat actors to easily configure phishing panels and modify existing templates using natural language commands, significantly lowering the technical barrier for deploying campaigns. The backend chatbot is model-agnostic and integrates with several frontier and open source platforms via APIs. The evolution of the Jieyou platform highlights a growing industry trend where PhaaS providers leverage LLMs as a modular force multiplier. Rather than manually editing payloads, low-skilled affiliates can now utilize natural language through the

Mitigations
These activities triggered Gemini's safety responses, and Google took additional, broader action to disrupt the threat actors' campaigns. Additionally, we've taken action against these actors by disabling the assets associated with this activity and making updates to prevent further misuse. Google DeepMind has used these insights to strengthen both classifiers and our models, enabling them to refuse to assist with these types of attacks.

How Google Protects Against AI Abuse

Google uses a multifaceted defense strategy to protect our users and infrastructure against LLM abuse, integrating proactive model-level safeguards, specialized threat intelligence, targeted containment protocols, and proactive red teaming to simulate and protect against threats. Additionally, as innovation moves forward, the industry needs security standards for building and deploying AI responsibly, which is why we introduced the Secure AI Framework (SAIF), a conceptual framework to secure AI systems.

We continuously harden our models against misuse by feeding insights from active threat monitoring directly into our safety classifiers and guardrails. For instance, in response to model extraction—or “distillation”—attacks, we have deployed real-time defenses designed to degrade the performance of unauthorized "student" models and detect attempts to clone proprietary logic. When we identify bad actors, we take direct action to disrupt their operations by disabling associated projects and accounts.

To extend these protections to enterprise customers, we developed Google AI Threat Defense (AITD). This autonomous architecture operationalizes security by bringing together the reasoning power of Gemini and other frontier models, the risk prioritization of Wiz, the automated remediation capabilities of Gemini and CodeMender, and frontline intelligence from Mandiant. AITD employs a multi-model strategy that balances cost and coverage, using light models for continuous scanning and specialized frontier models for high-risk vulnerabilities.

In addition to our proactive platform defenses, we’ve recently introduced Gemini 4 Carbon, our strongest model with state-of-the-art cybersecurity and long horizon capabilities. Google has built Gemini 4 Carbon with layered defenses against cyber misuse—combining safety training,

agentic trajectory monitors—designed to block requests to develop functional, weaponized exploits or malware tooling. The robustness of these safeguards has been stress-tested by both internal red teams and specialized external partners including Gray Swan and Trajectory labs. Beyond the model itself, Google continuously monitors for intentional misuse in production, limits or disables access for offending accounts, and provides vetted defenders, security researchers, and government partners access to advanced dual-use cyber capabilities through the Fairwind Program. More information on the safety, security, and alignment testing conducted on Gemini 4 Carbon can be found in the Gemini 4 Carbon Safety & Security Report.

Building AI Safely and Responsibly

Google's approach to AI is guided by a commitment to bold innovation and responsible development. Guided by our AI Principles, Google designs AI systems with robust security and safety guardrails, which are continuously tested to ensure resilience.

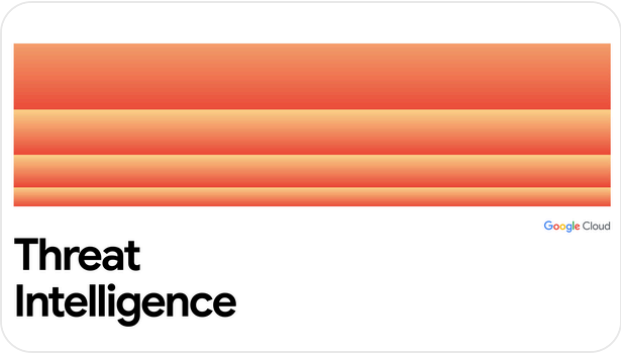
Our policy guidelines and prohibited use policies are foundational to ensuring safety. Our policy development process is built to anticipate emerging trends and design for security from the ground up, allowing us to enhance protections for users globally.

abuse of our platforms—including malicious cyber activities by government-backed threat actors—and collaborate with law enforcement when appropriate. Crucially, our learnings from every countermeasure we implement are fed back into our product development to improve the security for our AI models. These iterative improvements to our classifiers and model-level safeguards are vital to maintaining agility against evolving threats.

Our LLM development and Trust & Safety teams also work in constant concert with our threat intelligence, security, and modeling experts to effectively address misuse.

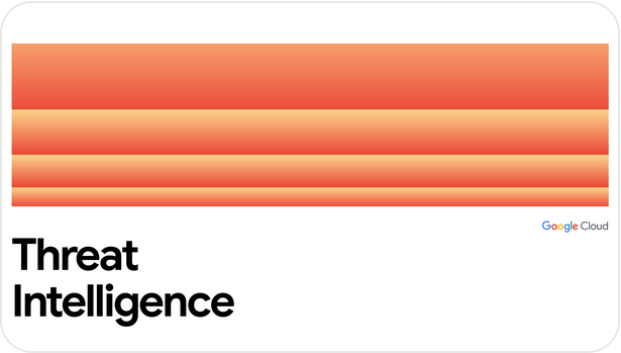
Posted in Threat Intelligence

Related articles



Threat Intelligence

ARMORY CASTLE Exploits vSphere



Threat Intelligence

COLD RELIC: A Case Study in Impersonation and Compromise



Threat Intelligence

The Rise of LLMJacking: How Threat Actors Monetize Compromised Cloud and Model Resources

By Google Threat Intelligence Group • 16-minute read



Threat Intelligence

Shutting the Door on the XCPE Botnet

By Google Threat Intelligence Group • 76-minute read