CSAI Foundation | Frontier Ready Cybersecurity

Frontier Ready Daily

Issue 82 · Member edition

2026-11-16



CSAI FOUNDATION INITIATIVE

Frontier Ready.

Machine-speed agentic cybersecurity

cloud security alliance®

CSAI

© 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.

In this issue

Google's threat intelligence group says it has, for the first time, seen attackers run patch-diffing and exploitation with no human in the loop, and its PaperCut timeline puts working LLM-built exploits within a day of disclosure and a mass campaign by day four. Amazon's decoy network recorded agentic campaigns harvesting AI provider keys from exposed model gateways and app builders, and the other items cover a criminal marketplace that uses AI to profile infostealer victims, Forescout's 90-hour sweep of ten embedded TCP/IP stacks, Anthropic's claim that Fable 5.2 blocks about 60% fewer legitimate security requests, and the FBI's ARTEX advisory. We examined Microsoft's MSRC post on risk cohorts for patch prioritisation (published in October, outside this window), Cloudflare's open-source Kitesurf agent browser (a product release with no security finding) and the Danfoss and NoisyTCP disclosures (individual flaws with no capacity effect), and left them out. No item today qualified under security_operating_model.

Google says attackers now diff patches and fire exploits with no human in the loop, and a PaperCut campaign hit hundreds of organisations four days after disclosure

Category: machine_speed **Significance:** major **Confidence:** VERBATIM (PROVIDER) for GTIG's statement that it identified, for the first time, actors using LLM frameworks to automate discovery and exploitation without a human in the loop, the exploit_agent framework that "automatically executes them without human intervention", zero-day exploits rising "from a 2025 average of 8 per month to 11 per month in 2026—before spiking to 22 in the month of August", and LLMjacking charges above $100,000 in a single day and up to $1 million cumulatively; LINK ONLY — VERIFY AT SOURCE for the PaperCut timeline (working LLM-built exploits within a day of disclosure, a GreyNoise-reported campaign against hundreds of organisations four days after disclosure, foothold to domain administrator "within minutes"), which GTIG attributes to Huntress and GreyNoise; CHARACTERIZATION (CSA) for treating one day as the planning assumption for internet-facing n-day exposure **Vendor interest:** Google sells Google SecOps, which GTIG says now carries ARTEX detection rulepacks, and Mandiant consulting, which has published a hardening guide for the same threat. **Source:** Google Cloud — GTIG AI Threat Tracker: Agentic Threat Activity Crosses a Significant Threshold

What changed. On 16 November 2026 Google Threat Intelligence Group reported a non-public multi-agent framework, found on an exposed staging server, that continuously compares popular web software before and after patches, drafts exploits and runs them without a person approving each step. GTIG calls

this the first fully automated discovery-to-exploitation pipeline it has observed. Its PaperCut example gives the interval: two chained zero-days were exploited in late August, researchers built working exploits with LLMs within a day of disclosure, and four days after disclosure GreyNoise saw an actor that likely used LLMs to build the exploit, assemble a target list and hit hundreds of organisations, reaching domain administrator within minutes of a foothold. GTIG also reports that zero-day exploits rose from 8 a month in 2025 to 11 a month in 2026, then 22 in August.

Why it reaches you. The component class is any internet-facing application with a public patch history: print management, web CMS, file transfer, edge appliances. Patch diffing used to need a skilled human and days of work, and that delay gave defenders their window. GTIG's account removes the human, and the PaperCut timeline shows the window is now a day or less. GTIG also notes a second cost: hundreds of non-working LLM-generated proofs of concept appeared after a PHP patch, and each one is triage work for a defender trying to decide whether it is exposed.

What to do. Vulnerability management owner, with the CISO: escalate. Set a written target for internet-facing n-days of mitigation (patch, virtual patch or removal from exposure) within 24 hours of a vendor fix, measure how often you meet it, and report the gap. Treat a public proof of concept as unverified until reproduced, so that fake exploits do not consume emergency capacity. Urgency: escalate.

CSA coverage. The Disclosure Bottleneck: When AI Discovery Outpaces Patch Consumption; see also Machine-Speed Attacks: Cloud Defense at the Inflection Point.

Amazon's decoys recorded agents working through exposed model gateways and AI app builders for provider keys, using default master keys and unauthenticated settings endpoints

Category: agentic_surface **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for the 3 October to 5 November MadPot collection window, the "hundreds of targets", the targeted products (One API, New API, LiteLLM, Dify, Langflow, Ollama and others), the moderate-confidence assessments that ARTEX was used and that the operators are separate China-nexus groups, the operators' rule that keys "exposed publicly for more than 30 days were treated as low-value inventory", and Amazon's statement that it "cannot verify the outcomes the agents reported"; LINK ONLY – VERIFY AT SOURCE for the technique list (LiteLLM default master key, unauthenticated provider-settings endpoints, the One API `isCliRequest` bypass, SSRF to cloud metadata) as reported by The Hacker News; CHARACTERIZATION (CSA) for the model-gateway-as-credential-vault reading **Source:** AWS Security Blog – Amazon Threat Intelligence observes threat actors running cybercrime campaigns with ARTEX AI agents

What changed. On 13 November 2026 Amazon Threat Intelligence published what agents sent to its MadPot decoys, which posed as popular LLM services, between 3 October and 5 November. The agents worked through hundreds of targets, prioritising LLM gateways and proxies and AI app builders because one weak admin page on a gateway can expose keys for many providers. Stolen keys were verified, pooled and fed into the framework's own model gateway to pay for further hunting. The operators treated keys exposed for more than 30 days as low value, which means they prize fresh exposures. Amazon could not see the attacks on real victims and did not verify the outcomes the agents reported, but it notified AWS customers who may have been affected. Separately, GTIG's report today names a LiteLLM flaw, CVE-2026-40217, that one LLMjacking actor used to turn limited virtual keys into full administrative credentials.

Why it reaches you. The component class is the model gateway and the self-hosted AI app platform. These are often deployed by product teams outside the normal hardening path, and they concentrate provider keys for many models in one place. The failure here is not a new technique but a deployed control missing: default master keys left in place, installs left unfinished, and management endpoints that need no login. The cost lands on your provider bill, and Mandiant reports single-day fraudulent charges above $100,000.

What to do. Platform security owner: validate. Inventory every model gateway and AI app builder reachable from the internet, confirm default admin credentials and master keys have been changed, confirm settings and provider-key endpoints require authentication, and patch LiteLLM, Langflow and RAGFlow. Put per-key spend alerts on every provider account those gateways hold. Urgency: validate.

CSA coverage. Five Minutes from Search to Key Vault: Agentic Secret Harvesting; see also Who Pays for Stolen Inference? Loss Allocation After Black October.

A criminal marketplace sells infostealer logs with AI-written profiles of each victim, searchable in plain language for corporate access such as Citrix

Category: agentic_surface **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for Hudson Rock's statement that the service markets "the context around the stolen data as part of the product", the Citrix VPN credentials in a clinic listing, the seller's claim of "valid Anthropic Claude API session tokens" in one log, and its advice to "revoke affected sessions and tokens"; SELF-REPORTED (PROVIDER METRIC) for the operator's claim of tens of millions of logs and more than a billion parsed accounts, which Hudson Rock has not verified (purchase records or an independent sample of listings would confirm it); LINK ONLY – VERIFY AT SOURCE for some matched logs coming from infections in the previous 24 hours, as reported by Help Net Security; CHARACTERIZATION (CSA) for the access-triage reading **Vendor**

interest: Hudson Rock sells infostealer-exposure intelligence, so the severity framing is commercially favourable to it. **Source:** Hudson Rock – Inside the AI-Powered Marketplace Turning Infected Computers Into Searchable Profiles

What changed. On 15 November 2026 Hudson Rock described a criminal marketplace that attaches an AI-generated profile to each infected machine it sells: who probably owns it, where they work, and which accounts are worth buying. Buyers can search in plain language, for example "citrix access to healthcare provider", and listings start at $0.25. Examples include a $9 machine in a defence contractor's household scored 100/100 and an AI researcher's machine that the seller says holds Claude API session tokens. Help Net Security reports that Hudson Rock matched several advertised logs to its own records, some from infections within the previous 24 hours.

Why it reaches you. The exposure path is an unmanaged or personal device that holds corporate sessions: VPN and Citrix gateways, SaaS cookies, and the configuration folders of coding agents. Sorting stolen logs for valuable corporate access used to take a skilled broker. With the sorting automated, a low-skill buyer can find your organisation's access by name within a day of infection. This feed reported on 9 November that infostealer operators signed in an average of eight minutes after infection, so the interval that matters is how fast you revoke sessions after an infection appears in a log.

What to do. Identity owner, with the SOC: validate. Measure the time from an infostealer infection on any device holding corporate sessions to revocation of those sessions and tokens, including VPN, SaaS and AI provider keys. Confirm whether you receive infostealer-exposure alerts for your domains at all and who acts on them. Urgency: validate.

CSA coverage. Living Off the Model API: Stolen Claude Keys as Agent Implants.

Forescout used commercial models to confirm 99 flaws in ten embedded TCP/IP stacks in about 90 hours, a job that took nine months of analysis by hand in 2020

Category: vuln_storm **Significance:** notable **Confidence:** SELF-REPORTED (PROVIDER METRIC) for 99 confirmed vulnerabilities (66 new), the 362 to 312 to 99 validation funnel, per-stack precision of 11% to 70%, about 10 hours of automated analysis plus about 80 hours of validation, and $1,200 to $2,300 in model spend (an independent re-run of the harness on the same stacks, or vendor confirmation of the 54 disclosed findings, would confirm the figures); VERBATIM (PROVIDER) for Siemens confirming affected APOGEE, Desigo and TALON controllers, as reported by Help Net Security, and for the defunct manufacturer of the access point with seven device-specific flaws; CHARACTERIZATION (CSA) for the unpatchable-backlog reading **Vendor interest:** Forescout sells device visibility and segmentation, so its advice that containing risk should take priority over patching is commercially favourable to it. **Source:** Forescout Vedere Labs – Finding Critical Infrastructure Bugs with Off-the-Shelf AI

What changed. In research published this month and reported by Help Net Security on 16 November 2026, Forescout said a harness built on Claude models and existing scanners confirmed 99 vulnerabilities across ten TCP/IP stacks used in industrial, medical and network equipment. Sixty-six were new, and 52% were rated high or critical. Its 2020–21 manual study of comparable stacks took about nine months of code analysis. This run took about 90 hours, most of it human validation, and $1,200 to $2,300 in model spend. The denominator matters: 362 candidates became 99 confirmed findings, and precision ranged from 11% to 70% per stack. One stack sat in a discontinued access point whose manufacturer no longer exists, so its seven device-specific flaws, including unchangeable administrator credentials, will not be patched.

Why it reaches you. The component class is embedded networking code in OT, building automation, medical devices and network gear, where patch cycles run in years and some vendors are gone. Discovery cost has fallen by roughly two orders of magnitude while remediation capacity for these devices has not changed, so the gap between known and fixed flaws will widen fastest in the estate you can patch least. The 80 hours of validation against 10 hours of discovery also shows where the bottleneck now sits for vendors receiving AI-generated reports.

What to do. OT and IoT asset owner: validate. Identify devices whose vendor is defunct or whose support has ended, and confirm each is isolated with its management services (web admin, Telnet, FTP) unreachable from user and internet segments. Add stated support lifetimes and a vulnerability disclosure policy to device procurement requirements. Urgency: validate.

CSA coverage. When Discovery Outruns Remediation: Open-Source Disclosure Strain; see also The Widening Gap: AI Discovery Outpaces Patch Capacity.

Anthropic says Fable 5.2 blocks about 60% fewer legitimate security requests than 5.1, but blocked requests are still silently answered by an older model

Category: defender_models **Significance:** context **Confidence:** SELF-REPORTED (PROVIDER METRIC) for "about 60% fewer" unnecessary blocks on cybersecurity requests in early testing across Claude.ai, Claude Code and Cowork, with no baseline rate, test size or threshold published (an independent over-refusal test on a fixed set of legitimate defensive tasks would confirm it); VERBATIM (PROVIDER) for the safeguards covering "penetration testing, exploit generation" and routing flagged requests to Opus 4.8, as reported by BleepingComputer; LINK ONLY – VERIFY AT SOURCE for the reporter's example of Fable 5.1 rerouting a request to log in to a server it had itself configured; LIVE TEST REQUIRED for over-refusal on your own defensive workflows **Vendor interest:** Anthropic sells the models described, and the refusal figure is its own measurement. **Source:** BleepingComputer – Anthropic's Claude Fable 5.2 blocks fewer security tasks, while Mythos gets better at finding bugs

What changed. BleepingComputer reported on 16 November 2026 that Anthropic claims Fable 5.2 cuts unnecessary blocks on cybersecurity requests by about 60% compared with Fable 5.1. When its safeguards flag a request as high-risk cyber activity, such as penetration testing or exploit generation, the request is answered by Opus 4.8 instead. Anthropic has not published the baseline refusal rate, the size of the test or where the classifier threshold sits. This feed covered the Mythos 5.2 containment figures on 10 November. The new point is the over-refusal claim for the general-access model.

Why it reaches you. The component is any defensive agent or analyst workflow built on a general-access frontier model. A defensive pipeline that is silently routed to a different, older model on some requests produces outputs of uneven quality with no visible failure. A relative reduction with no baseline tells you nothing about the absolute rate you will see.

What to do. Security engineering owner running AI-assisted triage or testing: monitor. Keep a fixed set of legitimate defensive tasks, run it against each model version you use, record refusals and reroutes, and log which model actually answered. Urgency: monitor.

CSA coverage. Cyber Defender Models: A CISO's Guide to Access, Testing, Containment, and Governance.

The FBI's ARTEX advisory says some intrusions reached privileged access in under three hours and lists the overlooked assets the framework favours

Category: machine_speed **Significance:** context **Confidence:** LINK ONLY — VERIFY AT SOURCE for the FBI FLASH's statements that multiple actors have used ARTEX since at least August 2026, that some intrusions reached privileged access "in under three hours", the sectors named, and the favoured targets (partner portals, employee applications, legacy services, third-party integrations), as reported by BleepingComputer; LINK ONLY — VERIFY AT SOURCE for the more than 1,600 exposed ARTEX consoles cited by researchers; CHARACTERIZATION (CSA) for the reading that the advisory consolidates earlier reports rather than adding new incidents **Source:** BleepingComputer — FBI warns ARTEX AI hacking tool is powering global cyberattacks

What changed. On 13 November 2026 the FBI issued a 13-page FLASH advisory with indicators of compromise for ARTEX, the open-source agentic pentest framework whose original repository was removed on 8 October but whose forks and a commercial edition remain available. The advisory says multiple actors have used it since at least August, that some intrusions reached privileged access in under three hours, and that it lets low-skill actors run operations against several targets at once. It consolidates earlier vendor reports rather than describing a new incident, and warns that its IP indicators may already be stale.

Why it reaches you. The FBI names the assets ARTEX finds easiest: partner portals, employee applications, legacy services and third-party integrations. These are the assets least likely to sit inside a tight patch and monitoring programme. A federal advisory also gives security leaders a citable basis for asking for rate limits and burst detection on these endpoints.

What to do. SOC owner: monitor. Load the advisory's indicators with an expiry date, and prioritise the behavioural detections it recommends, such as bursts of automated requests and unexpected privilege changes, over the IP lists. Urgency: monitor.

CSA coverage. ARTEX and Its Forks: Agentic Pentest Frameworks as Attack Infrastructure.

Watchlist

- **OpenAI reward-hacking postmortem — downstream response** — No change. No frontier lab disclosed an eval-to-production escape, and there was no new regulatory or contractual development. *(opened 2026-08-27)*
- **VM/hypervisor containment hardening for cyber-capable agents** — No change on microVM adoption for agent sandboxes. *(opened 2026-08-27)*
- **Claude Code Auto Mode prompt-injection ASR discrepancy** — No change. There is still no Anthropic patch or public response. GTIG reports that a suspected state actor modified `.claude/settings.json` in internal repositories so that starting Claude Code downloaded a backdoor, which is a configuration-tampering path rather than prompt injection. *(opened 2026-08-27)*
- **AI defensive-triage guardrail evasion** — No change. No new campaign triggering refusals to blind triage tooling was reported. Anthropic's claim of about 60% fewer unnecessary cyber blocks in Fable 5.2 (item 5) concerns over-refusal of legitimate work, not adversarial triggering. *(opened 2026-08-31)*
- **AI account session hijacking at scale** — Movement. Hudson Rock (item 3) documents a marketplace whose seller advertised a log with Claude API session tokens, and GTIG reports LLMjacking actors using session hijacking and token-capturing lookalike sites for initial access. Neither confirms a provider-scale campaign, and no provider has shipped device-bound session tokens. *(opened 2026-08-31)*
- **Disclosure deadlines shrink toward 30 days** — GTIG's PaperCut timeline (item 1), with working exploits within a day of disclosure and a mass campaign by day four, is a further data point for the zero-soak assumption. No vendor response yet to Project Zero's schedule. *(opened 2026-11-14)*

- **Agent session traces as a credential store** — No change. No further Fireworks AI victims named and no vendor moved training opt-out to organisation level. *(opened 2026-11-15)*