

AFTERNOON EDITION — WEDNESDAY 28 OCTOBER 2026

Reporting window: 27 Oct 2026, 16:20 CET to 28 Oct 2026, 16:10 CET (Paris time). Issued 28 Oct 2026, 16:10 CET.

This afternoon edition has **eight dispatches**: seven new to this journal and one update. Five were published today. The update adds **CNBC's confirmation** of OpenAI's plan to buy Google's chips, which this morning rested on one unnamed source. Two Anthropic research posts and Mozilla's Firefox 159 advisory are dated 27 October and were missed this morning, so the window starts at yesterday's 16:20 edition; the Firefox advisory appeared around that edition's close, and its OpenAI credits were first tallied today.

8 NEW OR UPDATED DISPATCHES · 0 BACKGROUND ITEMS · PREVIOUS: [2026-10-28-MORNING](#)

THE NEWS

MODELS **NEW**

NVIDIA RELEASES NEMOTRON 3.5 TRANSCRIBE, AN OPEN-WEIGHT SPEECH MODEL THAT TRANSCRIBES 40 LANGUAGES AND TELLS SPEAKERS APART

PUBLISHED: 28 Oct 2026, 15:58 CET

Nvidia on Wednesday released **Nemotron 3.5 Transcribe**, an open-weight model that transcribes **40 languages**, translates 31 of them to or from English and labels up to eight speakers in 13 languages. It handles recordings up to an hour long and can answer questions about the audio from the same checkpoint.

TechCrunch puts the model at **30 billion parameters**; Nvidia's notes say about 4 billion are active at a time and that it runs on a single Nvidia GPU, from DGX Spark to Blackwell systems. Nvidia says it ranks first among open-weight models on the Hugging Face Open ASR leaderboard's accuracy-for-size frontier, at a 5.33% average word error rate.

Nvidia also released **Granary v2**, an open speech dataset and curation pipeline, with fine-tuning recipes. In Nvidia's own test, fine-tuning cut the word error rate on held-out speakers by 45.8% relative to the best prompt. TechCrunch notes that the model works on recorded audio, not live streams.

ANALYSIS — WHY IT MATTERS. *Analysis: Transcription for meetings and medical dictation is one of AI's most established paid uses. An open model that also handles speakers and translation puts pressure on commercial speech APIs.*

CAVEAT. Leaderboard positions and error rates are **Nvidia's own figures**; language support differs by task, and audio question answering is trained in English.

SOURCES

- Hugging Face blog (Nvidia): Nemotron 3.5 Transcribe and Granary v2** <<https://huggingface.co/blog/nvidia/nemotron-transcribe-granary>> (primary source) **CHECKED: 28 Oct 2026, 16:04 CET**
- TechCrunch** <<https://techcrunch.com/2026/10/28/voice-ai-in-vogue-nvidia-releases-new-open-transcription-model/>> (reporting) **CHECKED: 28 Oct 2026, 16:04 CET**

CODING & AGENTS **NEW**

FIREFOX 159 FIXES 167 SECURITY FLAWS, AND MOZILLA CREDITS 41 OF THEM TO OPENAI'S CODEX SECURITY AGENT AND AN OPENAI RESEARCHER

PUBLISHED: 27 Oct 2026

Mozilla's advisory for **Firefox 159**, dated 27 October, lists **167 fixed vulnerabilities**, each with its own CVE. Mozilla credits **34 of them to OpenAI Codex Security**, OpenAI's security agent, and seven more to OpenAI researcher Amy Burnett.

Notebookcheck, which tallied the credits on Wednesday, counts 108 flaws rated high, 52 moderate and 7 low, none critical, and notes that the advisory reports no active exploitation. Most of the high-severity bugs are use-after-free flaws, in which a program keeps using memory it has already released. The only sandbox escape on the list was found by Mozilla itself.

Mozilla has changed how it reports bugs: it no longer groups internally found memory-safety bugs under a single CVE, so the total is not directly comparable with older releases.

ANALYSIS — WHY IT MATTERS. *Analysis: Alongside Anthropic's exploit report, this shows AI bug hunters feeding the patch pipelines of major browsers in volume. The harder question is whether fixing keeps pace with finding.*

CAVEAT. Mozilla's credits show **who reported each bug**, not how much human review went into the reports.

SOURCES

- Mozilla: Security Vulnerabilities fixed in Firefox 159 (MFSa 2026-118)** <<https://www.mozilla.org/en-US/security/advisories/mfsa2026-118/>> (primary source) **CHECKED: 28 Oct 2026, 16:05 CET**
- Notebookcheck** <<https://www.notebookcheck.net/Firefox-159-fixes-167-flaws-and-OpenAI-s-Codex-found-41-of-them.1436136.0.html>> (reporting) **CHECKED: 28 Oct 2026, 16:05 CET**

CISCO OPEN-SOURCES PROJECT QUARRY, A RULE LANGUAGE FOR SPOTTING HARMFUL BEHAVIOUR ACROSS WHOLE AI AGENT SESSIONS

PUBLISHED: 28 Oct 2026, 13:45 CET

Cisco on Wednesday open-sourced **Project Quarry**, a detection language and engine for AI agent behaviour. A security team describes a multi-step attack once and checks every agent session against it; Cisco likens it to what its Snort tool did for network traffic and YARA for malware, The Register reports.

A rule can combine **machine-learning classifiers** or **LLM judges** with exact checks, such as a string match or the order of two tool calls. Quarry covers conversations, tool use, Model Context Protocol traffic and images, and when a rule matches it returns the steps and excerpts behind the finding, with common secret formats masked.

The release includes a command-line scanner, a Python SDK, a starter rule pack and a plugin for Claude Code, Codex and other MCP-compatible agents. Cisco says Quarry already runs in its AI Defense product and that it plans to add it to DefenseClaw, its open-source governance layer for agent runtimes.

ANALYSIS — WHY IT MATTERS. *Analysis: Attacks on agents often look harmless one step at a time. A shared, testable rule format would let defenders trade detections the way they already share malware signatures.*

CAVEAT. Cisco's researchers say rules **do not replace human oversight** or chain-of-thought monitors; no outside measurement of detection rates has been published.

SOURCES

- Cisco Blogs: Introducing Project Quarry** [\(<https://blogs.cisco.com/ai/project-quarry-agent-security-detections>](https://blogs.cisco.com/ai/project-quarry-agent-security-detections) *(primary source)* **CHECKED: 28 Oct 2026, 16:05 CET**
- GitHub: cisco-ai-defense/project-quarry** [\(<https://github.com/cisco-ai-defense/project-quarry>](https://github.com/cisco-ai-defense/project-quarry) *(primary source)* **CHECKED: 28 Oct 2026, 16:05 CET**
- The Register** [\(<https://www.theregister.com/security/2026/10/28/cisco-open-sources-tool-for-detecting-ai-agents-on-hacking-sprees/5305867>](https://www.theregister.com/security/2026/10/28/cisco-open-sources-tool-for-detecting-ai-agents-on-hacking-sprees/5305867) *(reporting)* **CHECKED: 28 Oct 2026, 16:04 CET**

ANTHROPIC SAYS AN UNRELEASED CLAUDE MODEL WROTE WORKING EXPLOIT CHAINS AGAINST CHROME, SAFARI, ANDROID, WINDOWS AND LINUX

PUBLISHED: 27 Oct 2026

Anthropic's Frontier Red Team on 27 October described **five attack chains** built by an **unreleased internal research model** against some of the most widely used software. In each case, Anthropic says, Claude wrote the exploit after initial prompting.

The chains are a one-click remote code execution in Chrome that escapes the browser's sandbox; a one-click Safari attack that disables the same-origin policy to read data from other sites; extraction of photos from a locked Google Pixel 9 over USB; a remote crash of Windows servers through the kernel's QUIC code; and a **zero-click root exploit for Linux servers** with a multipath TCP listener. Anthropic says its Chrome reports have produced **62 CVEs**, **all rated critical or high**.

The flaws were reported under coordinated disclosure, which gives vendors 90 days to patch. Anthropic says Apple, Google and Microsoft have fixed the Safari, Android and Windows issues, and that the Linux bug was patched upstream. In its proof of concept, the Linux attack took 48 minutes to leak a kernel pointer and five more to run a root command.

Anthropic recommends scanning code when changes are proposed, not only afterwards, and has released a Claude Code security plugin for that purpose.

ANALYSIS — WHY IT MATTERS. *Analysis: Chaining several bugs into a working exploit has long been the work of a few specialist teams. If models can do it with little guidance, the time defenders have to patch shrinks, which is why Anthropic is steering its strongest cyber tools to vetted security professionals.*

CAVEAT. The model is **not named or released**, and the account is Anthropic's own; the prompts and the amount of human help were not published.

COMMUNITY REACTION (OPINION, NOT VERIFIED FACT)

- “doesn't it make sense to at least run the browser in isolated environment nowadays?”
— [cromka](#) ([Hacker News](#))
- “But can they fix and harden the software without introducing new bugs in the meantime?”
— [pllbnk](#) ([Hacker News](#))

SOURCES

- Anthropic: Discovering and exploiting vulnerabilities in critical, real-world software with Claude** [\(<https://www.anthropic.com/research/discovering-vulnerabilities-with-claude>](https://www.anthropic.com/research/discovering-vulnerabilities-with-claude) *(primary source)* **CHECKED: 28 Oct 2026, 16:04 CET**
- Hacker News discussion (27 Oct)** [\(<https://news.ycombinator.com/item?id=50265205>](https://news.ycombinator.com/item?id=50265205) *(community)* **CHECKED: 28 Oct 2026, 16:06 CET**
- Hacker News discussion (28 Oct)** [\(<https://news.ycombinator.com/item?id=50270131>](https://news.ycombinator.com/item?id=50270131) *(community)* **CHECKED: 28 Oct 2026, 16:06 CET**

ANTHROPIC STUDY: ANALYSTS WENT ALONG WITH AN AI AGENT'S SECRETLY ASSIGNED LEANINGS 86% OF THE TIME

PUBLISHED: 27 Oct 2026

In a randomized trial Anthropic published on 27 October, 50 professional analysts used Claude Code to advise a fictional farming cooperative on whether to expand an advisory programme. Unknown to them, a coin flip set which of two defensible options their agent would argue for on each of four analytical choices. They went with the agent's lean 86% of the time, against about 50% expected by chance.

Analysts who had already stated a preference switched away from it 60% of the time when the agent disagreed, against 6% when it agreed, an estimate that rests on only 38 cases. Afterwards, 77% justified their choices with the agent's arguments, sometimes word for word, but only 15% credited the agent. A statistical model using only the four coin flips predicted 82% of the final recommendations.

The agents ran Claude Opus 4.8 and were instructed to argue honestly, to change their minds if the evidence demanded it and never to mislead. Of 61 analysts recruited, 11 chose not to use the agent and were left out of the main analysis.

ANALYSIS — WHY IT MATTERS. *Analysis: Keeping a human in the loop assumes the human is an independent check. The study suggests that on open judgment calls an agent's framing can quietly become the expert's own, which Anthropic notes also applies to AI safety research run with agents.*

CAVEAT. Anthropic deliberately strengthened the agents' leanings; it studied one hour-long task, in one field, with one model, so effects in real work may be smaller.

SOURCES

- Anthropic: How AI agents shape experts' judgment calls**
<<https://www.anthropic.com/research/agents-judgment-calls>> (primary source) **CHECKED:** 28 Oct 2026, 16:04 CET
- Anthropic paper (PDF): Who's Aligned to Whom?**
<<https://www-cdn.anthropic.com/519c17a5d23a794e8b6d50a854048d0b61e0d73a.pdf>> (primary source) **CHECKED:** 28 Oct 2026, 16:07 CET

UPDATE: CNBC CONFIRMS OPENAI PLANS TO BUY GOOGLE'S TPU CHIPS, AS GOOGLE TURNS TO PARTNERS TO SELL MORE OF THEM

PUBLISHED: 28 Oct 2026, 12:00 CET

This morning's edition reported that OpenAI's plan to buy Google's AI chips rested on one unnamed source cited by The Information. On Wednesday CNBC said it had confirmed that OpenAI is planning to buy tensor processing units (TPUs) from Google. CNBC gave no size, price or timing for the purchase.

CNBC presents Oracle's deal as a change of strategy for Google, which began selling TPUs to outside customers and data centres only in the second quarter of this year; Oracle's cloud adds another sales channel. Analyst estimates vary widely: TD Cowen models \$62.5 billion of TPU sales in 2027, and Wells Fargo expects about \$99 billion next year.

Anthropic remains Google's biggest TPU customer, according to CNBC. TD Cowen estimates a 20% operating margin on external TPU sales, below Google Cloud's overall margin of about 35% last quarter.

ANALYSIS — WHY IT MATTERS. *Analysis: If the purchase goes ahead, Google's two biggest model rivals would both run on its chips, adding pressure on Nvidia. Alphabet reports quarterly results on Thursday.*

CAVEAT. CNBC does not describe how it confirmed the plan, and no company statement on the purchase was found.

SOURCES

- CNBC** <<https://www.cnbc.com/2026/10/28/google-alters-cloud-playbook-turning-to-partners-to-sell-more-ai-chips.html>> (reporting) **CHECKED:** 28 Oct 2026, 16:05 CET

MALAYSIA SEIZES 30 SERVERS OF ADVANCED AI CHIPS AT KUALA LUMPUR AIRPORT, ITS FIFTH CASE AGAINST ONE SMUGGLING NETWORK THIS YEAR

PUBLISHED: 28 Oct 2026, 06:52 CET

Malaysian customs said on Wednesday that officers had seized 30 server units containing advanced AI chips, worth about 49.5 million ringgit (US\$12.1 million), in the free-trade zone of Kuala Lumpur International Airport on 15 October.

The servers were declared as electronic components and were to be re-exported from one Asian country to another without the required permit, airport customs director Aspalila Ag Tuah said. It is the fifth case linked to the same syndicate this year, with seizures worth more than 100 million ringgit in total; no arrests have been made.

Malaysia has required permits for the export and transit of advanced US-origin AI chips since last year, after US pressure to stem their flow to China. Officials did not name the chip brands, origin or destination, citing the ongoing investigation.

ANALYSIS — WHY IT MATTERS. *Analysis: Repeated seizures show that transit hubs remain a route around chip export controls, and that enforcement still catches shipments case by case.*

CAVEAT. Customs did not say where the chips were headed; the link to China is policy background, not a finding in this case.

SOURCES

- CNA (Reuters)** <<https://www.channelnewsasia.com/asia/malaysia-airport-customs-seize-advanced-ai-chips-smuggling-6498976>> (reporting) **CHECKED: 28 Oct 2026, 16:05 CET**
- Malay Mail (Bernama)** <<https://www.malaymail.com/news/malaysia/2026/10/28/klia-customs-seizes-rm495m-in-ai-chips-with-potential-military-uses-fifth-case-in-rm100m-syndicate-probe/240981>> (reporting) **CHECKED: 28 Oct 2026, 16:05 CET**

OPENAI SETS UP AN ELECTION OPERATIONS CENTRE TO WATCH FOR MISUSE OF ITS TOOLS DURING THE US MIDTERMS

PUBLISHED: 28 Oct 2026, 15:00 CET

OpenAI is standing up an Election Operations Center, a 'war room' that will run from this week until all votes are counted in next week's US midterm elections, Politico reported on Wednesday as an exclusive. Politico says the effort brings together hundreds of employees.

The centre is meant largely to catch abuse of OpenAI's own tools, such as ChatGPT. It will monitor events in real time and handle reports of unexpected model behaviour or suspected breaches of the company's terms.

Meta said on 27 October that it had activated its own election operations centre and begun its usual ban on new political ads for the campaign's final week.

ANALYSIS — WHY IT MATTERS. *Analysis: AI companies are being drawn into the election-integrity role that social networks have played for a decade, while their own models are among the tools being watched.*

CAVEAT. The details come from one outlet's exclusive; no OpenAI post describing the centre was found.

SOURCES

- Politico** <<https://www.politico.com/news/2026/10/28/openai-election-war-room-interference-01137690>> (reporting) **CHECKED: 28 Oct 2026, 16:04 CET**
- Politico Decoded newsletter** <<https://www.politico.com/newsletters/decoded-by-politico/2026/10/28/silicon-valleys-historic-wealth-tax-crusade-01137428>> (reporting) **CHECKED: 28 Oct 2026, 16:04 CET**
- Meta: How Meta is preparing for the 2026 US midterm elections** <<https://about.fb.com/news/2026/02/meta-prepares-for-2026-us-midterms/>> (primary source) **CHECKED: 28 Oct 2026, 16:04 CET**

KEY PATTERNS — EDITORIAL ANALYSIS

- AI bug hunting has reached volume: five Anthropic exploit chains and 34 Codex Security credits in one Firefox release.
- Human oversight of agents is under test: analysts adopted agents' assigned leanings, and Cisco offers rules to watch whole agent sessions.
- Chip supply keeps spreading and leaking: OpenAI's TPU plan is confirmed, while Malaysia seizes another batch of smuggled servers.

SOURCE COVERAGE & COLOPHON

COLLECTION ENGINE

VERSION: 3.28.0
RAN THIS EDITION: yes
ENGINE LOOK-BACK: 2 days

SOURCE STATUS

REDDIT: ok (15 threads)
YOUTUBE: ok (5 videos, 2 with transcripts; some transcript requests hit HTTP 429)
HACKERNEWS: ok (21 stories)
POLYMARKET: unreachable (TLS certificate hostname mismatch; not bypassed)
GITHUB: ok (12 items)
DEVTO: searched with tags ai and llm; 0 articles returned
X: not searched (no consent-free source configured)

COVERAGE NOTES

- The window runs from Tuesday afternoon's 16:20 CET issue to this edition, overlapping this morning's window to catch three items it missed. Dispatches with only a date carry that date; exact timestamps are given where a source supplied one.

ENGINE FOOTER

All agents reported back!
└─ Reddit: 15 threads | 3,111 upvotes | 1,656 comments
└─ YouTube: 5 videos | 3,112,647 views | 2/5 with transcripts

- The installed last30days engine finished with 53 evidence items from Reddit, Hacker News, GitHub and YouTube. Its top items were general-audience videos on an 'AI bubble' and AI-produced proofs, and threads on stories already reported, such as GPT-6.1 Terra and the Artificial Analysis index. Hacker News threads on Anthropic's exploit report supplied the community quotes. Stories were chosen with Google News RSS, publisher feeds, web searches and checks of primary documents. Item counts are not counts of verified developments.
- Polymarket failed its TLS certificate hostname check and verification was not bypassed. X was not searched. dev.to returned no articles. Some YouTube transcript requests hit HTTP 429. Failed or missing coverage does not mean nobody discussed these stories.
- The configured web search backend failed and its keyless fallback served the searches. Politico (HTTP 403) refused automated access, so the OpenAI election dispatch relies on Politico's text as returned by web search; The Washington Post did not respond, VentureBeat's feed returned HTTP 429 and Google DeepMind's feed could not be parsed.
- Not carried: Cathedral's AI-agent intrusion into Anthropic's business systems (Washington Post, 26 October; scope disputed), Microsoft's voluntary-exit offer (26 October), HHS's HTI-5 rule dropping the AI 'model card' requirement for certified health IT (released 26 October), Mistral's Responses API preview (no launch date found), President Trump's 'Super Intelligence Force' (announced 4 October), Amazon's guidance to engineers on AI token use (no primary source checked) and Senator Warren's letter to Anthropic (WSJ, paywalled).
- No browser-cookie access, optional service or credential signup, paid enrichment, TLS bypass, Job Alerts edits or schedule changes were used. Research logs and full provenance remain private.

THIS EDITION

- [Exact input JSON](#)
- [Markdown text](#)
- [Permanent page](#)

- HN: 21 stories | 1,207 points | 818 comments
- GitHub: 12 items | 7 reactions | 20 comments
- Top voices: r/artificial, r/singularity, r/LocalLLaMA
-

The AI Gazette is a locally rendered digest. Each dispatch separates reported facts, labeled analysis and caveats, and cites the sources checked at the times shown. Background items are older or undated context, not new reporting.

Rendered as static HTML with inline styles and no scripts, cookies or analytics. Typefaces are requested from Google Fonts (fonts.googleapis.com, fonts.gstatic.com); if that request is blocked, Georgia, Times or another installed serif is used.