

AI models found critical holes in the PHP core

Attacks on AI keys at Railway and LiteLLM, the Claude Duet agent, Mistral Large 4 and a \$25 billion Meta penalty

Thursday, 22 October 2026

IN THIS ISSUE

1. PHP2Shell: models found critical holes in the PHP core, and a leak broke the release
2. Railway and LiteLLM: two attacks in one day, both for keys to AI
3. Anthropic: Claude found 493 unnecessary entry points in the code of 25 projects
4. Anthropic showed Claude Duet, an agent that writes first
5. Mistral Large 4: one trillion parameters, weights before the end of the month
6. Meta: record revenue, negative cash flow and a \$25 billion penalty
7. Muse from Meta: one query found 6,800 "private" links
8. OpenAI models will come to Google Cloud
9. Thousands of GPT-Rosalind agents examined a mouse brain atlas in 32 minutes

TOPIC 1

PHP2Shell: models found critical holes in the PHP core, and a leak broke the release

SOURCES The PHP Foundation, 21.10 [Thephp](#) · Wiz Research, 21.10 [Wiz](#) · OpenAI, 21.10 [OpenAI](#) · Elias Taylor, 21.10 [Eliastaylor](#) · Simon Willison, 21.10 [Simon Willison](#)

Yesterday this digest wrote that PHP prepared a critical update. On 21.10 it came out: PHP 8.5.12, 8.4.27, 8.3.36 and 8.2.35. The update closes two vulnerabilities, CVE-2026-109618 and CVE-2026-109619, both with CVSS 9.2. Both sit in memory management and in the handler of fatal errors. A request that hits `memory_limit` at a bad moment leaves a corrupt internal state. The next request in the same worker takes this state as valid. Thus an attacker without a login gets either a read of process memory or code execution.

Wiz writes that all versions back to PHP 5.4 are vulnerable. The researchers showed code execution without authentication on fresh images of WordPress, Drupal and Joomla. Separately, they showed a read of worker memory with passwords, tokens and API keys. Wiz worked with Wiz Atlas, Gemini 4 Argon and GPT-5.6 Sol. According to Wiz, the models made the exploit practical. The exploit needs a long chain of steps for the memory layout of each target.

The story of the fix is longer than the story of the discovery. The PHP Foundation writes that one of its developers found the first signs of the problem in 2024, during an audit by Quarkslab. The security team could not select one of several fix options, and the work stopped for months. The foundation directly calls this a structural failure. On 30.07, PHP joined the OpenAI program Patch the Planet, and the firm Calif started an audit. On 11.08, Calif and OpenAI independently reported code execution with the same root cause on the same day. Wiz sent its report on 24.08. Since 01.07, the PHP team published 15 CVEs. It also examined more than 100 reports that were false, duplicates or not reproducible.

The coordinated disclosure broke two days before the date. On 20.10 at 15:00 UTC, the CVE numbers appeared online with a wrong release time and a wrong CVSS score of 10.0. PHP moved the release from 22.10 to 21.10 at 14:00 UTC. After that, Calif published a video with a working exploit by mistake.

Calif then deleted it, but copies spread. On 21.10 at 07:41 UTC, Chainguard broke the embargo and published built images with the fix. Thus anyone could study the patch in binary form six hours before the release.

The threat estimates differ. OpenAI calls this one of the most important security problems of recent years. It promises \$1 billion of subsidized access to Daybreak for defenders. Willison writes that exploits circulated since the night. At the time of the release, the PHP Foundation saw no working exploits. The developer Elias Taylor checked public PoCs from GitHub. He writes that they do not work on clean PHP. Most of them need an existing hole in the application, namely a local file inclusion.

Why it matters. Models found a vulnerability that lived in the core for more than ten years.

Three parties confirm this independently: the PHP Foundation, Wiz and OpenAI. The second half of the story is more important for any maintainer team. The flow of reports from AI tools already exceeds the capacity of small teams. And an embargo breaks because of people and processes. The practical conclusion: update on the day of the release.

TOPIC 2

Railway and LiteLLM: two attacks in one day, both for keys to AI

SOURCES Railway, 21.10 [Railway](#) · SafeDep, 21.10 [Safedep](#) · @levelsio, 21.10 [@levelsio](#)

On the morning of 21.10, the cloud platform Railway saw suspicious activity in the systems that keep customer environment variables. According to Railway, an outsider got access to encrypted variables and decrypted part of them. Such variables usually contain API keys and database credentials.

Railway tells all customers to change all variables and registry keys. All deploys on the platform are on hold until the investigation ends. The company thinks that the attack is part of a series of strikes on platforms. The goal of the series is keys to inference providers. Railway asked the providers to revoke all keys.

On the same day, SafeDep examined the LiteLLM 1.107.0 release on PyPI. Many teams use this library as a gateway to different LLMs. All seven wheels of the release carried a password stealer and a backdoor. The malicious code started with any Python process on the machine, even without an import of litellm. The stealer collected SSH keys, cloud credentials and CI secrets. It also collected tokens of Claude Code, Cursor, Codex and other AI tools. It also took LLM provider keys from the LiteLLM proxy database. Each hour, the backdoor read a server address from a file on GitHub and ran code from there.

The GitHub repository has no malicious code. SafeDep checked 539 GitHub Actions runs and did not find the run that built the release. Its version, with medium or high confidence: a private BerriAI pipeline built the wheels from substituted code. The GitHub account with the server address disappeared at 13:40 UTC, approximately two hours after the release. PyPI removed the release.

Versions 1.106.2, 1.107.0rc1 and 1.107.0rc2 are clean.

Why it matters. A key to an AI provider became a separate target. It gives money, access to data and compute. Nobody confirmed a link between the two attacks, but they have the same target. If LiteLLM 1.107.0 got onto at least one machine, treat the keys from that machine as compromised. A reinstall of the package does not help here. Rotate all secrets.

TOPIC 3

Anthropic: Claude found 493 unnecessary entry points in the code of 25 projects

SOURCES Anthropic, 20.10 [Anthropic](#)

Anthropic described a different approach to code protection. The model looks for code that is reachable without need. Then it puts this code behind authentication or explicit activation. Claude went through 25 open projects, more than 5 million lines in C, C++, Java, PHP and Python. It made a list of more than 17,000 configuration options, and 3,842 of them relate to security. In total, the model found 493 problems.

Each of the 11 projects that must work only after a login had code that was reachable before the login. There, the model found 93 unnecessary pre-authentication surfaces in the default configuration. The system accepted 162 more setting combinations without an error and without documentation. In 94 of them, the combination silently disabled protection. In one case, two options together disabled the check of the client certificate. Then a server that the admin configured for trusted clients only accepted anyone.

One pattern appeared in four independent projects. An administrator writes a deny rule by host name.

Reverse DNS is off, thus the server compares the rule with the IP address as text. A host name never matches an IP, and the deny list lets everyone through.

Why it matters. A search for individual vulnerabilities scales badly: a fixed hole tells nothing about the rest. Less reachable code also reduces the weight of holes that nobody found yet. This is research by the company itself, and there is no independent check of the numbers yet. But any administrator can check access rules by host name today.

TOPIC 4

Anthropic showed Claude Duet, an agent that writes first

SOURCES Anthropic, 21.10 [Anthropic](#) · The New York Times, 21.10 [NYT](#) · [@TheRundownAI](#), 21.10 [@TheRundownAI](#)

Claude Duet is a separate app for iOS and Android. On a computer, it opens on [claude.ai](#). The agent runs on Opus 5.5. It does not wait for a request. Each morning, it sends a summary from mail, calendar and work apps. Then each few hours it checks these accounts and writes when something needs attention. Users give it tasks in messages, as text or voice. The agent works on its own cloud computer and has access to more than 6,000 connectors and plugins.

Before it sends a message or makes a payment, the agent asks for consent. The NYT writes that it can buy through Link from Stripe and Shop Pay from Shopify, but a human approves each purchase. Anthropic blocked stock trades and transfers between accounts, even with consent. A separate filter stops the agent from the transfer of card numbers and document numbers.

Duet is available as a research preview with a waitlist, only for Pro and Max in the US. There is no separate fee, but the background checks use the subscription limits. Anthropic itself writes that the agent can misread a thread or make an error about urgency. The NYT puts it next to dots from OpenAI, Muse from Meta and Gemini Spark from Google. Unlike them, Anthropic bets on work, and the agent does worse with household tasks.

Why it matters. A chatbot waits for a human, and the Duet agent comes to the human itself. Thus the main question now is how often the agent interrupts and what it does without oversight. Watch how the background checks consume the limits, because a usual subscription pays for them.

TOPIC 5

Mistral Large 4: one trillion parameters, weights before the end of the month

SOURCES Mistral AI, 21.10 [Mistral](#)

Mistral opened a public preview of Mistral Large 4 through the API. It is a multimodal model with 1 trillion parameters, and 52 billion of them are active. The company promises to publish the weights before the end of October. Until then, the model goes through red teams with cybersecurity partners and government bodies. They get a version with weaker moderation and extended cyber capabilities.

Mistral trained the model from scratch on 3,800 NVIDIA Grace Blackwell GPUs in its own data centres in Europe. The company itself gives the numbers. On DeepSWE v1.1 the model scores 61.7%, on Terminal-Bench 4 it scores 28.3%. In a blind code evaluation by Surge AI, it is second of five models: 3.74 points against 4.22 for Claude Opus 5. In the Artificial Analysis Cyber Index test, the model reproduces a real vulnerability and patches it. There, Mistral claims 82%, the highest result among all models.

Why it matters. Mistral directly sells open weights as an argument for cybersecurity. A refusal by a provider in the middle of an incident is also a risk, and an own model removes it. There are no independent measurements yet, and "weights before the end of the month" is only a promise for now.

The check comes when the weights appear on Hugging Face.

TOPIC 6

Meta: record revenue, negative cash flow and a \$25 billion penalty

SOURCES The New York Times, 21.10 [NYT](#) · Semafor, 21.10 [Semafor](#) · CNBC, 21.10 [Cnbc](#) · New Mexico Department of Justice, 21.10 [Nmdoj](#)

Meta reported for the third quarter. Revenue grew by 26%, to \$64.6 billion. Costs grew by 68%, to \$51.6 billion. Profit was \$9.8 billion. Meta raised the lower limit of annual capital spend from \$130 to \$135 billion, and most of it goes to data centres. Free cash flow for the quarter became negative: minus \$2.12 billion. Long-term debt grew to \$80.9 billion.

To get money for AI, Meta will issue a new class of shares in January and do a 2-for-1 split. Semafor writes that the new shares will have no voting rights. Thus the company can sell shares, pay engineers and buy companies without a dilution of Zuckerberg's control. Alphabet made a similar move in 2014.

On the same day, a court in New Mexico gave Meta a \$25 billion penalty in a case that grew from the Cambridge Analytica scandal. In September, the jury counted more than 43 million violations of the unfair practices law, up to \$5,000 for each. The state got the right to \$219.5 billion, but the judge reduced the sum to \$25 billion. Interest on the sum accrues at a rate of 15% per year.

Why it matters. Even a company with record ad revenue no longer covers AI spend with money from its business. Big tech companies stopped share buybacks and started to sell shares. For an investor, this is a dilution of the stake. For the market, it is one more signal of the price of the compute race.

TOPIC 7

Muse from Meta: one query found 6,800 "private" links

SOURCES Ars Technica, 20.10 [Ars Technica](#) · The New York Times, 21.10 [NYT](#)

The Muse agent lets users share a conversation through a link. All such links live on subdomains of `cf.metaaiusercontent.com`, and the subdomain name contains the conversation title. When a recipient opens the link, the DNS query goes into commercial passive DNS databases. Researchers from the firm Claroc got 6,800 links with one query to such a database.

They opened a few hundred of them. They found an internal revenue review for a CEO, a loan application, notes for a confidential interview with an AI lab and the vacation schedule of a full company. Claroc reported the problem through the bug bounty program on 08.10 and got no reply. After a question from Ars, Meta paused the share function. According to the NYT, Muse has more than 10.5 million downloads since 08.09.

Why it matters. A long random string in a link gives no protection if the address gets into DNS.

Any service with links on subdomains has the same weakness. Keep the secret of a link in the path or in a parameter. The host name is not a safe place for it.

TOPIC 8

OpenAI models will come to Google Cloud

SOURCES OpenAI, 21.10 [OpenAI](#) · Ars Technica, 21.10 [Ars Technica](#)

OpenAI and Google Cloud agreed to add OpenAI models and Codex agent workflows to Model Garden.

Customers will buy them through Google Cloud Marketplace. Companies with GCP spend commitments will count the purchases against these commitments. Later, OpenAI models will appear in Google Antigravity and Gemini Agent. The companies promise general availability in the next weeks.

Separately, OpenAI writes that Microsoft stays its main cloud partner.

Why it matters. At large companies, the cloud budget is often already allocated. An option to pay for OpenAI from the Google budget removes a procurement barrier. Competitors of Gemini now sell in the store of Google itself. The selection of a model for each task becomes a usual practice.

TOPIC 9

Thousands of GPT-Rosalind agents examined a mouse brain atlas in 32 minutes

SOURCES OpenAI, 21.10 [OpenAI](#) · @OpenAI, 21.10 [@OpenAI](#)

On 16.10 this digest wrote that OpenAI opened GPT-Rosalind for academic labs. Now OpenAI, PerturbAI and Scripps Research showed the first large work with it. Thousands of agents examined a CRISPR atlas of the full mouse brain in 32 minutes. Each agent took one

gene knockout and analyzed the response of the cells. It also searched the literature and suggested an explanation. The team estimates such a volume of manual work at several months or a year.

According to OpenAI, the agents reproduced the conclusions of humans and found new hypotheses. One links cholesterol transport with neuron degeneration through the gene DCTN1. Another explains why TSC2 mutations cause a more severe form of tuberous sclerosis than TSC1 mutations. In a blind comparison, the novelty scores from the agents matched the scores of experts.

Why it matters. A scheme of thousands of parallel agents gives hypotheses. A lab must still test each of them. There is no peer-reviewed paper yet, and the company gives all the numbers. The value is in the speed of the first pass over large data, which otherwise stays unexamined.

misc

- **Tesla and SpaceX:** Musk confirmed that the special committees of both companies discuss a merger. This continues item 4 of 21.10. He says that the deal needs the consent of at least 50% of Tesla shareholders. On the same day, Tesla reported a 43% drop in quarterly profit.
[BBC · Ars Technica](#)
- **eMudhra:** mitmproxy added the option `--emudhra`, which signs certificates with the compromised intermediate key of eMudhra. This continues item 3 of 21.10. In the same commit, mitmproxy removed the eMudhra authorities from its builds.
[GitHub](#)
- **GitLab** released a critical update for all branches from 18.11 to 19.5. The most severe vulnerability, CVE-2026-110264, sits in the regular expression compiler and affects all versions from 7.0. [Gitlab](#)
- **@TZahavy (Google DeepMind):** a pipeline on Gemini 4 Argon found 20 new "cryptid" candidates for the BB(6) problem. These are tiny Turing machines whose halt is equivalent to open problems of number theory. The proofs are formalized in Lean 4. [one source]
[@TZahavy](#)
- **Qwen3-Coder:** the firm Hirundo writes that the model gives a vulnerable app 19% more often when the customer is a US defense contractor. Hirundo sells a "westernization" service for models, thus this is research by an interested party. [one source]
[Hirundo](#)
- **Google DeepMind:** the BBC spoke with several employees who left because of the military contracts of the company. Google answers that it checks each contract against its AI principles.
[one source] [BBC](#)

- **DuckDB 2.0.0 LTS** came out with asynchronous I/O, the VARIANT type, triggers and a mode for agents. Support will last at least until October 2027.

[Duckdb](#)

- **@ycombinator**: Legora raised \$550 million at a valuation of \$8.8 billion. More than 140,000 lawyers work on the platform together with agents.

[@ycombinator](#)

- **@AnthropicAI**: the physicist Christopher Stubbs, former dean of science at Harvard, joined the Long-Term Benefit Trust of the company. [@AnthropicAI](#)