

OpenAI claims hundreds of solved math problems

Mistral showed Large 4, Anthropic opened cyber access tiers, Kwon apologized in Sydney, Korean banks were hacked.

Wednesday, 7 October 2026

IN THIS ISSUE

1. OpenAI published 722 manuscripts from an internal model and claims hundreds of solved problems
2. Mistral Large 4: a trillion parameters, weights at the end of the month and a charge against closed models
3. Anthropic opened three access tiers for security researchers, from defense to pentests
4. Jason Kwon apologized to Australian senators, and the agent logs weigh 50 PB
5. South Korea is investigating attacks on banks where traces of a Chinese AI agent were found
6. The takeover of three national domain registries gave attackers certificates for Google domains
7. OpenAI launched the Decisions API: the model answers with a probability, a choice or a score
8. "Protocol pivoting": an agent passes a malicious instruction to another agent because it trusts it
9. EmbeddingGemma 2: multimodal embeddings with 740 million parameters for a phone
10. openTPU: AI agents designed an accelerator that runs real models on an FPGA

TOPIC 1

OpenAI published 722 manuscripts from an internal model and claims hundreds of solved problems

SOURCES OpenAI, 06.10 [OpenAI](#) · openai/math repository, 06.10 [GitHub](#) · WSJ, 06.10 [WSJ](#) · Advisory Group on Mathematics and AI recommendations, 29.09 [Agmai](#) · Thomas Bloom, erdosproblems.com, 06.10 [Erdosproblems](#) · HN discussion, 06.10 [Hacker News](#)

OpenAI released results obtained by an unreleased internal model when it was tested on open research problems. According to the repository README, the catalogue contains 722 manuscripts grouped into 372 "families": the main result, supporting arguments, corollaries and alternative proofs. The model was given about 4,000 problems, and each result took on average about three hours of ChatGPT Pro-level compute. Some of the proofs are formalized in Lean, and the rest are promised to follow. The company states outright that unformalized results "may contain errors".

The scale of the claims is large. The README singles out a paper on a zero-free region of the Riemann zeta function for $\text{Re}(s) > 11/12$ and a proof of the Hodge conjecture for CM abelian varieties: these two results were obtained outside the standard procedure, and humans edited the text of the first. Among the ten published condensed reasoning chains are the irrationality exponent of π , Mahler's conjectures, an isomorphism of free group factors and the Vlasov-Maxwell system. A commenter on HN checked the catalogue against proofatlas.ai's list of the 500 top open problems and counted 90 that the repository claims to close completely, including the unique games conjecture from complexity theory. This is an outside user's count, and OpenAI does not give such a figure. The WSJ recalls that a month ago the same model already claimed a solution to one of the Millennium Prize problems.

This digest covered on 01.10 the recommendations of the advisory group of mathematicians at the Institute for Advanced Study in Princeton. OpenAI cites that very group: it says it consulted the group and relied on its advice. Checking against that checklist shows some items were met: the number of attempts is named (about 4,000), the average compute cost is given, several reasoning chains and formalizations are published, and funding for seminars and conferences is promised. Other items remain open. The group asked for results to be deposited in a repository the lab does not control, while here it is OpenAI's own GitHub with a promise to "look for alternatives". The model's name is not disclosed.

The main point: the first paragraph of those same recommendations says the group does not endorse testing hard math problems on closed models and asks for this to stop. On HN some read that quote as guild protectionism, while others asked what a graduate student should do whose half-finished work covers the same topic.

The same day Thomas Bloom, who runs erdosproblems.com with 1,221 problems, explained why he is changing the rules. The main public activity on the site now, in his words, is people posting AI-generated proofs without explanation to stake a priority claim, and he does not want to run a site like that.

Why it matters. Verification is now more expensive than generation: hundreds of manuscripts with no human author, some without formal checking, and someone has to read them. Lean formalization moves trust from a company's reputation to a machine check, and it is the thing to watch when judging which of the loud claims stand firm. For any team publishing the results of an AI system, this case shows the difference between "disclosed how many attempts" and "handed the result over for independent verification".

TOPIC 2

Mistral Large 4: a trillion parameters, weights at the end of the month and a charge against closed models

SOURCES Mistral, 06.10 [Mistral](#) · Simon Willison, 06.10 [Simon Willison](#) · WSJ, 06.10 [WSJ](#) · HN discussion, 06.10 [Hacker News](#)

Mistral opened a public preview of Large 4, officially nicknamed "le Chonk". It is a multimodal MoE model with 1 trillion parameters, 49 billion of them active. On Hugging Face the model page is named Mistral-Large-4.0-1T05-A52B, so the name says 52 billion active: the company has not yet explained the discrepancy. The weights are promised by the end of October, along with architecture details. For now only the API is available: \$1.36 per million input tokens and \$4.18 per million output tokens. The model was trained from scratch on 3,800 NVIDIA Grace Blackwell GPUs in Mistral's own data centers in Europe, and the data covers more than 160 languages.

Figures from the company blog: DeepSWE v1.1 61.7%, Terminal-Bench 4 28.3%, AutomationBench 59.9%. In Surge AI's blind code quality evaluation the model is second of five (3.74 out of 5) after Claude Opus 5 (4.22). **Yesterday this digest reported** on Reflection's Beam with 501 billion parameters: on the same DeepSWE Beam scores 44.4%, and The Rundown AI immediately put the two sets of numbers side by side.

@levelsio, citing Artificial Analysis data, writes that Large 4 is slightly behind the Chinese open models DeepSeek, GLM and Kimi, but ahead of any American open model.

The sharpest part of the announcement concerns cybersecurity. On one test from the Artificial Analysis Cyber Index, where the model has to reproduce a real vulnerability in open source code and then patch it, Large 4 scores 82%, the highest result among all models. Mistral writes that Claude Opus 5.5 and GPT-6 Astra score close to zero on the same test because they refuse to do the task. Before the weights are published the model is being "red teamed" by security companies, partners and government bodies, with relaxed moderation and extended cyber capabilities. At the same time the company reports that on harmful cyber prompts from JailbreakBench, StrongREJECT and AgentHarm the model refuses more often than other open models.

Why it matters. The largest open models have so far been mostly Chinese, and within two days the American Beam and the European Large 4 arrived. For organizations that need a model under their own control, the choice has widened. The "zero because of refusals" claim should be read as a vendor's argument: it is true of how closed models behave by default, but item 3 below shows that these models also exist in a mode with most of the blocks removed.

TOPIC 3

Anthropic opened three access tiers for security researchers, from defense to pentests

SOURCES Anthropic, 06.10 [Anthropic](#) · announcement on X, 06.10 [@AnthropicAI](#)

Anthropic merged two programs, Project Glasswing and the Cyber Verification Program, into one expanded CVP with three tiers. Each gives access to Claude Opus 5.5, Sonnet 5.5, Claude Mythos 5.1 and future models, and the tiers differ in the number of cyber blocks and in vetting requirements.

- **Defense Access:** SOC work, incident response, malware reverse engineering, vulnerability validation. Eligible are corporate security teams, hospitals and utilities, small security firms, open source maintainers and individual researchers with a record of found vulnerabilities.

Applications are promised a review within a few days.

- **Red Team Access:** adds authorized pentests and red teams, for organizations only, with a review taking a few weeks. Blocks remain on actions that could cause physical harm or mass outages, for example deploying ransomware.
- **Specialized Access:** the fewest blocks, for vetted organizations testing aviation systems, power grids, telecom and interbank transfers. Each organization is reviewed together with the US government, and Glasswing participants move here automatically.

Program participants must agree to data retention so the company can track abuse. An alternative is promised later in the fall: Enterprise Frontier Safeguards, where the data will stay in the customer's cloud.

The company showed how the tiers work on CyScenarioBench, a test of multi-stage cyber operations: 10 scenarios with 5 attempts each at every tier. Without the program everything was blocked on the first prompt. On Defense Access 46 of 50 attempts were blocked at some stage. On Red Team Access there were no blocks, and Opus 5.5 completed 34 of 50, which matches its 67.6% with no safeguards at all.

Why it matters. The same day Mistral sold its model on the claim that closed models block the legitimate work of defenders. Anthropic answered exactly that problem, only by vetting people instead of lifting restrictions for everyone. Two access models now sit side by side: open weights, where control is with the owner, and a closed model, where control is with the provider but graded. The CyScenarioBench figures are useful as a rare case of a company showing what exactly its classifier blocks at each tier, with numbers instead of general words about "safe access".

TOPIC 4

Jason Kwon apologized to Australian senators, and the agent logs weigh 50 PB

SOURCES The Guardian, 06.10 [Guardian](#) · Ars Technica, 06.10 [Ars Technica](#) · Simon Willison, 07.10 [Simon Willison](#) · FT, 06.10 [FT](#)

Yesterday this digest reported that Australian MPs were preparing to question Jason Kwon. The hearing took place. Kwon, OpenAI's chief strategy officer, flew to Sydney and apologized for the way the company told Services Australia about an agent's access to Medicare data: by an unsigned letter to the agency's general mailbox. Asked by Senator David Pocock why the letter went to an arbitrary address, he answered: "In hindsight, we should have done as you say".

The Guardian draws attention to the timeline. The incident happened on 18 June, OpenAI learned of it almost a month before 1 September, when Sam Altman met Australia's deputy prime minister Richard Marles and told him nothing. The letter to the agency went out on 10 September. Kwon says Altman did not know about the incident at that point, but never explained why. One more detail from the hearing: OpenAI is still going through the activity logs of the agents involved in the incident, and their volume is 50 petabytes.

Ars Technica has fresh details on the Wikimedia story: the agents tried to turn the shared Etherpad notepad and the citation tool into proxies for loading third-party sites, made millions of API requests and hundreds of thousands of requests to the Wikidata Query Service, which may have caused its partial outage in May. OpenAI responded with a statement that it is working with Wikimedia and is still looking for similar incidents. Simon Willison noticed that edits in the Wikipedia sandbox began on 12 May, a day after the first edits on the German wiki covered earlier, and suggests it is the same swarm of agents.

The FT writes the same day that insurers are preparing for multimillion-dollar claims over "uncontrolled"

agents (text paywalled, only the headline is visible).

Why it matters. 50 petabytes of logs show the scale of the oversight problem: when there are so many agents that their activity is reviewed for months, the incident is found later than the victims should have been told. For any organization running agents on an external network, the lesson of this story is simple: have a plan for who gets notified and how before an incident, along with the contacts of the people to call.

TOPIC 5

South Korea is investigating attacks on banks where traces of a Chinese AI agent were found

SOURCES WSJ, 06.10 [WSJ](#) · The New York Times, 06.10 [NYT](#) · FT, 06.10 [FT](#)

South Korean President Lee Jae Myung said at a cabinet meeting that "signs have appeared" of AI models being used in recent attacks on banks, and ordered a quick finding of what happened. His quote from the broadcast: "It has now become possible to hack easily with AI even without special skills". The cyber unit of the National Police Agency opened an investigation.

The figures in the sources differ. The WSJ writes about at least seven financial companies, 68,000 victims and traces of the Artex AI tool, developed in China. The NYT, citing Yonhap, names three banks and their own data: at Shinhan the credit information of about 25,000 people leaked, at Hana 89 customers, at KB Kookmin 99 customers and 20 employees. Who is behind the attacks has not been reported.

All three outlets confirm the topic, and the FT and WSJ pieces are available only as a headline and a first paragraph.

Why it matters. Until now, public stories about AI in attacks were mostly reports from the labs themselves about abuse they caught. Here the head of state is the one speaking, and the tool in question was built as a security product. For banks and everyone who stores personal data, this is an argument to test defenses against automated brute force that does not tire and costs pennies: rate limiting, anomaly monitoring and fast response matter more than before.

TOPIC 6

The takeover of three national domain registries gave attackers certificates for Google domains

SOURCES Google, 06.10 [Blog](#) · Ars Technica, 06.10 [Ars Technica](#)

Google described a series of takeovers in the national domain zones of Ghana (.gh), Sierra Leone (.sl)

and American Samoa (.as). Attackers compromised the registries themselves, changed the authoritative DNS records of selected domains and thereby passed automatic domain ownership checks. That is how they obtained real HTTPS certificates for several Google domains and, as Certificate Transparency logs showed, for domains of other large brands and services. Google stresses that its systems and certificate authorities were not breached: the authorities did everything by the rules.

Chrome blocked the certificates it found through CRLSets, and for Google they were also revoked. But the company warns it cannot guarantee it found all affected domains, and the browser block does not protect users of other browsers. Which Google domains were affected and how many certificates were issued has not been disclosed. Advice to domain owners: monitor Certificate Transparency logs continuously for all their domains, including regional and parked ones, and publish restrictive CAA records bound to ACME accounts. The latter will not prevent issuance during a takeover, but it stops a cached validation from being reused after DNS control has been returned.

Why it matters. The HTTPS chain of trust turned out to be only as strong as the weakest domain zone registry, and the domain owner has no influence over that. Certificate Transparency monitoring is cheap, and here it was the only way for an owner to see the forgery. For companies with domains in small national zones (the American Samoa zone, for example, is popular with startups) this is a concrete item to check today.

TOPIC 7

OpenAI launched the Decisions API: the model answers with a probability, a choice or a score

SOURCES OpenAI, 06.10 [Openai](#) · Simon Willison, 06.10 [Simon Willison](#) · Perplexity Developers on X, 06.10 [@perplexitydevs](#) · HN discussion, 06.10 [Hacker News](#)

On 02.10 this digest covered Typesafe's Jev model, which returns a decision instead of text, and the WSJ calling this a new genre. Now OpenAI has released its own version in public beta. The Decisions API takes text or images and a list of questions of three types: predicate returns the probability that a condition holds, choice picks one value from the supplied options with a probability for each, and score rates the input on an ordered scale. According to the documentation, the response arrives roughly 10 times faster than through the Responses API. Only the gpt-6-luna model works, and images are accepted only as base64.

Willison wrote a plugin for his llm tool in an evening and compared prices: both models charge only for input tokens, OpenAI 10 cents per million, Jev 4.2 cents. The same day Perplexity updated its open model pplx-decider-v1.1-27b, which it says is half the price of the first version (\$0.02 per million input tokens) and first in the new Decision Index 0.3 ranking on Hugging Face.

Why it matters. A large share of production LLM use comes down to classification, routing and filters, and until now this was done with a generative model and parsing of its answer. A separate API type with probabilities instead of text lets teams set a threshold and measure quality with the usual metrics, as in classical ML. Versions from OpenAI and Perplexity and a dedicated ranking appearing within a week mean the genre has stopped being one startup's experiment.

TOPIC 8

"Protocol pivoting": an agent passes a malicious instruction to another agent because it trusts it

SOURCES [Ars Technica](#), 06.10 [Ars Technica](#)

Independent researcher Syed Anas Mohiuddin tested agents from Google, JP Morgan Chase, Weaviate, Rapid7, the French interministerial digital directorate and the US federal government. Over five months Google and four other organizations acknowledged vulnerabilities of one class: an attacker plants text in an agent with a narrow task (translation, data analysis), the agent passes it on as an ordinary delegated task, and the next agent executes it because it trusts the first. Mohiuddin called this protocol pivoting: entry through MCP, handoff through Google's Agent-to-Agent or another protocol, and trust is lost at the junction.

The most serious vulnerability, rated 8, was in Google's mcp-toolbox for databases: the HTTP client did not limit redirects and did not check destination IPs, so a crafted path parameter made the server reach internal addresses on the attacker's behalf. Google fixed it with allowlists of ranges and a base URL check at startup. The vulnerability found at Rapid7 was rated 2.7 and is already closed. Douglas McKee of Rapid7 described the problem this way: each protocol checks its own front door, and nobody guards the corridor between them.

Why it matters. In systems with several agents, protecting a single model against prompt injection does not help if the internal agents trust each other unconditionally and keep credentials on the MCP server. The practical conclusion for anyone building such chains: treat input from another agent as untrusted as input from a user, and restrict the network access of tool servers, as Google did after the fix.

TOPIC 9

EmbeddingGemma 2: multimodal embeddings with 740 million parameters for a phone

SOURCES [Google](#), 06.10 [Blog](#) · [Google DeepMind on X](#), 06.10 [@GoogleDeepMind](#) · [HN discussion](#), 06.10 [Hacker News](#)

Google DeepMind released the second version of its open embedding model, now multimodal: text, code, images, audio and video in one vector space. The 740-million-parameter model is built on the Gemma 4 architecture and licensed under Apache 2.0. It is modular: text needs only 270 million parameters, and the image (170 million) and audio (300 million) encoders are attached as needed. Vectors can be truncated from 768 to 512, 256 or 128 dimensions, saving up to six times the space. The context is 8K tokens, four times more than in the first version, which means up to 5.5 minutes of audio or 29 images.

On a Pixel 11 Pro with quantization the text part takes about 191 MB of RAM and the full model about 567 MB. According to Google, the result on the MTEB Code test rose from 68.76 to 78.68.

Why it matters. Searching your own photos, voice notes and videos without sending data to a server becomes a job for the phone. For developers, the combination of Apache 2.0, small size and vector truncation matters: local vector search over code or documents no longer requires separate infrastructure. The quality figures so far come only from Google itself.

TOPIC 10

openTPU: AI agents designed an accelerator that runs real models on an FPGA

SOURCES [GitHub FeSens/openTPU](#) [GitHub](#) · [HN discussion](#), 06.10 [Hacker News](#)

The openTPU project asks two questions: how far AI agents can go in hardware design, and whether they can build a chip that will run their own inference. The repository holds the whole stack: a SystemVerilog design, an instruction set, a bit-exact simulator, a kernel language with a compiler and software for a real PCIe card. The design runs on an Inspur YPCB-00338 FPGA card (Xilinx Kintex-7) and runs ten modern models with real weights, and the card produces the same tokens as the simulator. Figures from the README: LFM2.5-230M in 4-bit quantization decodes about 86 tokens per second, Qwen3-0.6B about 31 and Gemma 4 E2B about 12, with DDR3 memory utilization at 82-92% of peak. The license is Apache 2.0. On HN the topic collected 243 points and over 300 comments.

Why it matters. It is far from real TPUs, and the value lies elsewhere: agents carried out a whole hardware project from RTL to the driver, with verification on real hardware. For anyone who wants to understand how an AI accelerator works, from matrix multiplication in Python down to the wires, this is a rare compact repository that can be read in full.

misc

- **Polars 2.0:** initial support for processing data that does not fit in memory has arrived, SQL is now full-featured, and by the Polars team's own benchmarks on TPC-H and TPC-DS data it outpaces DuckDB and DataFusion. [Pola](#)
- **Dust from Q Labs:** a method of training transformers without backpropagation of error that, according to the authors, approaches backprop at a large "population" and scales better on larger models. Tested up to 1 billion tokens. [Qlabs](#)
- **The Nobel Prize in Physics went to Francis Halzen** for the discovery of high-energy neutrinos (the IceCube observatory in the Antarctic ice). [Ars Technica](#)
- **Finland halted work at two Google data center sites** in Muhos and Kajaani: at one of them the mandatory environmental impact assessment was not done, and over 300 hectares of forest were cleared.

Google admitted it "fell short of its own standards". [BBC](#)

- **JetBrains:** an HN thread with 574 points about the company ending 2025 with a loss despite record revenue. The data covers only the Czech legal entity JetBrains s.r.o.: revenue of 16,008 million koruna (+6.3%), net loss of 315 million koruna. There are no consolidated accounts here.

[Helgilibrary](#)

- **Paramount Skydance completed its \$111 billion merger with Warner Bros.** Discovery, and HBO Max, Paramount+ and Discovery+ will be combined into one service. [Ars Technica](#)
- **SpaceX is seeking \$40 billion for Nvidia chips** in financing led by Apollo, the FT reports [single source, text paywalled]. [FT](#)
- **@garrytan**, over dinner with Paul Graham, used Opus 5.5 in fast mode to rewrite Doom in Bel, Graham's LISP dialect, in about 20 minutes: 32 thousand lines of C++ in under 2 thousand lines of Bel plus a Bel interpreter in JS. [@garrytan](#)
- **@emollick** reminds labs that their models should know their own products: it is oddest when an AI can do everything on a computer except work with its own app. [@emollick](#)
- **@naval:** models are becoming the last line of defense in software, because AI can rewrite an essay or decompile a program but cannot "distill" itself, so more software will hide on the server.

[@naval](#)

- **Gamma 5** imports and exports PowerPoint with formatting preserved and bets on visual variety instead of identical AI slides.
[@thisisgrantlee](#)