AI DIGEST

OpenAI shows the perplexity of its models

Dragos finds breaches at inference providers, 87% of Anthropic power to come from gas, Replit tests agent teams.

Monday, 23 November 2026

IN THIS ISSUE

1. OpenAI published the perplexity of its models and proposes shared limits for the industry
2. 2 million trap prompts: Dragos found breaches at three providers, OpenRouter disconnected three suppliers
3. Better Data Center Project: in 2028 gas will give 87% of the energy of Anthropic, coal 11%
4. The Nvidia story: a senator demands to question Bessent and Lutnick, Bessent does not exclude a Huang return
5. Replit: 32 Opus agents made NumPy 18.9 times faster, GPT-6.2 teams did worse than one agent
6. The Delta outage in Atlanta and an explanation of the wave of outages in the US
7. GL.iNet: an attacker used a hole in the cloud to put malware on routers and KVMs [single source]
8. South Korea: the education minister resigned after a scandal with AI cheating [single source]

TOPIC 1

OpenAI published the perplexity of its models and proposes shared limits for the industry

SOURCES OpenAI, 22.11 OpenAI · @merettm, 22.11 @merettm · @gdb, 22.11 @gdb · HN Hacker News

Jakub Pachocki, the chief scientist of OpenAI, published the essay "Perplexity and Intelligence". Perplexity shows how well a model predicts the next token. The lower it is, the better the model compresses data. For the first time, OpenAI showed a chart of this metric for its base models, from GPT-3.5 to an internal model that is not released yet. Pachocki admits that these numbers give competitors a clear picture. But in his words, shared safety rules are more important now.

The numbers from the essay are as follows. GPT-3.5 had a log-perplexity of about 0.99 and could hold a conversation.

GPT-4 had about 0.91 and already gave useful chain-of-thought reasoning. Below 0.8, "transformative" abilities appeared in cybersecurity and mathematics. On a model with a value of about 0.76, OpenAI started to get breakthrough results on open math problems.

Pachocki expects the next window of new abilities at 0.68-0.7. He does not know what exactly will appear there. The measurement uses a closed set of internal OpenAI code.

The sharpest part of the essay is about speed. The chart shows a sharp acceleration in 2026. OpenAI explains it mostly with long research and compute, and only to a smaller degree with automation. But Pachocki writes that already in 2027 AI can give an "avalanche" of fundamental improvements to pre-training itself. In his opinion, progress in alignment and monitoring will not keep the same speed. Thus he proposes a shared limit: labs do not go past a certain perplexity level until they reach a specific goal in the understanding of models. One example of such a goal is a successor to chain-of-thought monitors.

The essay itself names the weak points of the metric. Labs measure on different datasets. Perplexity does not count reasoning at answer time. And nobody knows its lower limit. Greg Brockman supported the text and called it a base for shared safety thresholds. The HN thread got 519 points.

Why it matters: until now, labs agreed on safety through tests of specific abilities. Such tests become old fast, and a model learns to recognize them. Pachocki proposes to measure the "raw material" of intelligence itself, before any fine-tuning. The proposal will work only if other labs publish their numbers on comparable datasets. So far, no lab did this.

TOPIC 2

2 million trap prompts: Dragos found breaches at three providers, OpenRouter disconnected three suppliers

SOURCES Dragos, 19.11 Dragos · OpenRouter, 22.11 Openrouter · HN Hacker News · HN Hacker News

HD Moore, the author of Metasploit and now head of research at Dragos, sent about 2 million prompts to AI services from July to November. Each prompt had unique traps: a URL, a DNS name and a real AWS key from an empty account. Any access to a trap means that someone read the prompt. The report came out on 19.11. On 22.11, OpenRouter answered with its own announcement.

There are three findings. First: attackers broke into three inference providers and copied requests at the entry point.

In one case, they stole several hundred AWS keys, and somebody tested the keys within a few hours. Then the attacker searched for data in S3 and Secrets Manager and tried to enable models in Bedrock. None of the three providers knew about the breach until Dragos told them. Second: sellers on an inference marketplace offered a model at the lowest price, enabled request logging and read the prompts. One person on a home internet connection in France tried passwords from the prompts for a root login over SSH.

The third finding is about privacy. A Chinese lab took secret URLs from more than half of the prompts to its models.

Among them were about two thirds of the requests with the Zero Data Retention (ZDR) flag. The same thing happened when a US provider sold the model of this lab. 24 days after the request, one link appeared in a Feishu work chat, and six devices in Beijing and Shanghai opened it.

OpenRouter writes that it found no breach of OpenRouter itself. The company disconnected three suppliers and removed 125 endpoints from ZDR status. Atlas Cloud, Novita, Phala and Together already published public reports. Together writes about a prompt leak for June and July.

Why it matters: a privacy flag works only at the first step of the route, and a request often goes through several routers. Dragos advises three steps: choose the provider explicitly, do not use marketplaces with the cheapest price and put your own traps in the context of agents. An agent sends into the prompt everything that it sees, keys included.

TOPIC 3

Better Data Center Project: in 2028 gas will give 87% of the energy of Anthropic, coal 11%

SOURCES Better Data Center Project, 16.11 Betterdatacenterproject · Blood in the Machine, 22.11 Bloodinthemachine · HN Hacker News

The report came out on 16.11, but in the window it became the second HN story of the day with 914 points and 915 comments. The Better Data Center Project is a nonprofit group that studies the energy of data centers. It analyzed the contracts of one company, Anthropic. As of October, Anthropic had contracts for 15 GW of compute. Of this, 5.9-9.6 GW have links to specific sites.

The main estimate: in 2028, gas will give about 87% of the energy of Anthropic, and coal 11%. The authors calculate which power plants will cover the new load on the grids. Without contracts for new renewable sources, these are gas and coal plants. Anthropic has public contracts for renewable energy only in Norway and Canada. At the same time, the company spent $3.4 billion on gas generators with a capacity of 1.7 GW.

The group estimates CO2 emissions in 2028 at 17-38 million tonnes, and the middle scenario gives 32 million tonnes. This is more than all electricity generation of California in 2024 (31.1 million tonnes). Simple turbines and engines make up 80% of the on-site gas generation, and they are dirtier than combined cycle plants. Separately, the report mentions Colossus 2 from SpaceX. Anthropic takes compute there since May, and most turbines there run without permits.

The authors themselves name the limits of the estimate. Formally, part of the emissions will go into the reports of landlords, for example Amazon or SpaceX. The emissions of Google, Microsoft and Meta also grow. Thus the comparison does not mean that Anthropic will become the largest polluter.

Why it matters: the report shows the mechanism. The cheapest available generation covers new load on the grid, and this is gas and coal. A promise to "expand the energy mix" changes nothing without specific contracts. The same logic applies to any company that builds a data center near a gas grid.

TOPIC 4

The Nvidia story: a senator demands to question Bessent and Lutnick, Bessent does not exclude a Huang return

SOURCES The Guardian, 22.11 Guardian · Semafor, 22.11 Semafor · Financial Times, 22.11 FT · @GavinSBaker, 22.11 @GavinSBaker

Yesterday the digest wrote about the Alice Huang network and $3.6 billion of Chinese money. On Sunday the story moved to US television shows. Republican senator John Kennedy on Meet the Press called the Huang affair "the most expensive sex since Helen of Troy". But Huang himself interests him less. Kennedy wants to question Commerce Secretary Howard Lutnick and Treasury Secretary Scott Bessent under oath. The question: how China got around export control.

The Guardian reminds readers of the context. Megaspeed, where Alice Huang was the chief executive, made chip sales of about $12 billion around export control. Critics say that Lutnick ignored warnings about smuggling. Bessent and others wanted to put Megaspeed on a blacklist, but the Commerce Department delayed this decision. Kennedy says that the heads of the department are either incompetent or "deliberately allowed it".

Director of National Intelligence Jay Clayton said that intelligence checks Alice Huang. He called Nvidia chips "the jewels that everybody wants". Bessent on Face the Nation said that the government spoke with the interim leadership of Nvidia and with lawyers about who was responsible for compliance. FT came out with a headline that Bessent does not exclude a Huang return. Investor Gavin Baker quoted the words of Bessent: "I said we need to look for someone from the outside. Jensen is now from the outside".

Why it matters: the question went beyond the Nvidia board and now concerns the US government. If the Senate calls Lutnick, the work of export control will become the subject of the hearings. A Huang return will then depend on more than the shareholders.

TOPIC 5

Replit: 32 Opus agents made NumPy 18.9 times faster, GPT-6.2 teams did worse than one agent

SOURCES Replit, 19.11 Replit · @vsachar, 22.11 @vsachar

Replit researchers released Optimization-Bench. An agent gets the code of a popular library and two hours to make a real workload faster without a change in the result. A separate machine does the check, and the agents have no access to it.

It also catches tuning to the tests. Replit compared one agent and teams of 8, 16 and 32 agents for six models. The analysis came out on 19.11, and it spread in the X lists on 22.11.

On NumPy, one Claude Opus 5.5 agent gave a 3.66x speedup, and a team of 32 agents gave 18.92x. Opus and Sonnet 5.5 teams always beat their solo agents, with a median of 1.30x on 16 and 32 agents. GPT-6.2 Astra was the strongest solo agent:

on NumPy it was more than three times faster than solo Opus. But its teams became worse as they grew. At 32 agents, they lost to its own solo result on three of four tasks.

The organization made the difference. Opus sent 82% of messages to a shared board, with a thread for each workload, and announced "touching btree.c" before an edit. The Opus coordinator almost never wrote to the workers. In GPT-6 Astra and GPT-6.1 Sol, the workers wrote to the coordinator directly, 20-22 messages each, and the coordinator became the bottleneck. On Git, all workloads share the same 15 files, and there the Opus teams did not beat solo.

Why it matters: a strong solo agent does not yet mean a strong team. Clear ownership of parts of the work and a shared log that anyone can read give the result. The same rules work in human teams.

TOPIC 6

The Delta outage in Atlanta and an explanation of the wave of outages in the US

SOURCES The New York Times, 22.11 NYT · Financial Times, 22.11 FT · The Guardian, 22.11 Guardian

On 18.11 the digest wrote that hurried patches after AI findings break systems in the US. On Sunday, at the start of Thanksgiving week, Delta flights through Atlanta had delays. The outage came from the system that calculates the weight and balance of an aircraft. According to Delta, about 1,158 of the 1,654 delayed flights had a link to this system. In the afternoon, Delta restored the system. The FAA stressed that its air traffic control systems had no part in this.

The Guardian analyzed the full wave of outages since 12.11. The conclusion is more prosaic than the rumors about a cyberattack: it is mostly coincidence. A new router caused the nationwide ground stop of flights. The FAA installed it as part of the replacement of copper lines with fiber, and this project runs since 2002. A bug from a July update caused the Microsoft outage. But the National Security Agency (NSA) told Congress that fast fixes of AI-found vulnerabilities caused its outage.

Why it matters: two causes act together. Old infrastructure gets no investment. At the same time, AI vulnerability search forces companies to update twice as fast. Each hurried update of an old system carries a risk of an outage.

TOPIC 7

GL.iNet: an attacker used a hole in the cloud to put malware on routers and KVMs [single source]

SOURCES GL.iNet, 20.11 Gl-inet · HN Hacker News

GL.iNet makes popular travel routers and KVM devices. On 20.11 the company found a vulnerability in its remote management service. An outside attacker used it and put malware on part of the routers and KVMs. The company closed the hole in the cloud, so there are no new infections by this path. But the fix in the cloud does not remove malware from devices that are already infected.

GL.iNet advises to install the fresh security firmware for each model, even without signs of infection. A sign of infection is the process `crond -i /tmp/.log`. The firmware removes the malware without a reset of the settings. The company does not say how many devices the attack affected.

Why it matters: cloud management is an entry point to a device that sits inside the network. A KVM gives full access to the keyboard and screen of the connected computer. If remote management is not necessary, disable it.

TOPIC 8

South Korea: the education minister resigned after a scandal with AI cheating [single source]

SOURCES The Guardian, 22.11 Guardian

It all started with an exam for tax advisers. The top candidate admitted that he took photos of the questions with a hidden micro camera, sent them to an AI system and read the answers from a smartphone. This happened four days before the suneung, the main university entrance exam of the country. The government turned off public wifi near the exam centers, doubled the number of supervisors and allowed only pencils. Staff covered the buttons on the clothes of candidates with stickers. The government sent an emergency alert to all phones in the country to find hidden devices.

The government promised at least $75,000 for information about cheating.

The exam went without large incidents. But the education minister Choi Kyo-jin, 72 years old, resigned. Earlier, at a hearing, he said that because of his age he was not sure that he could catch violators.

Why it matters: exams that check knowledge without access to information now require an almost military regime. A micro camera and a smartphone with AI cost less than any supervision. Thus the format of such exams will most probably have to change.

misc

- **The Masullo proof in Lean:** the community keeps a public map of the formalization of the manuscript of Alessandro Masullo. One participant on HN writes that hundreds of found errors do not yet mean a hole in the main idea, and the authors already fixed part of them. According to the comments, section 13 of the manuscript is almost empty.

  Hacker News
- **P vs NP live:** pequalnp.com shows a fleet of AI agents that attacks the P vs NP problem. On the morning of 23.11, the agents checked 371 of 553 lemmas. This is an experiment, and there is no proof. Pequalnp
- **Leonida:** a repository for the static recompilation of the Xbox version of GTA VI for PC. The authors converted all 26,503 shaders to SPIR-V, but the game does not start yet. Yesterday @amasad wrote about this project as an example of how a swarm of agents coordinates through GitHub issues. GitHub
- **Backblaze:** from 01.12, a new privacy policy allows the company to sell or share personal data for targeted advertising. According to the company FAQ, backup files are not part of this. Neowin
- **Autonomous weapons:** the Red Cross warns that soon it will be too late to regulate autonomous weapons. The CCW conference in Geneva agreed only to discuss a regulation instrument for the next two years. Guardian
- **The US and China:** Trump said that the US will work with China on "joint guardrails" for AI. Semafor links the change of tone to protests against data centers and to the results of the midterm elections.

  Semafor