

Meta reaches the frontier with Muse Wonder

Anthropic against distillation, an attack on supermarket fridges, the Codex CLI background daemon and GLM-Legal

Tuesday, 13 October 2026

IN THIS ISSUE

1. Meta reaches the level of the leaders with Muse Wonder
2. Anthropic: defenses against distillation cut the theft of Claude reasoning by 90-99%
3. Hackers heated the fridges in hundreds of French supermarkets to 50 °C
4. Codex CLI silently starts a background daemon that updates itself
5. UK minister: an AI agent breach of government systems is "likely, if not inevitable"
6. Zhipu released GLM-Legal for legal work
7. Epic Games moves its joint game with Disney to AI development
8. Russia was not ready for the data center war

TOPIC 1

Meta reaches the level of the leaders with Muse Wonder

SOURCES Meta, blog, 12.10 [Meta](#) · Artificial Analysis, 12.10 [Artificialanalysis](#) · The New York Times, 12.10 [NYT](#) · Financial Times, 12.10 [FT](#) · Alexandr Wang, X, 12.10 [@alexandr_wang](#) · Tobi Lutke, X, 12.10 [@tobi](#) · Ethan Mollick, X, 12.10 [@emollick](#) · HN discussion, 12.10 [Hacker News](#)

Meta released two models: Muse Wonder and Muse Wonder Full. The company trained them in its gigawatt-scale Prometheus data center. Muse Wonder is available now through the Meta Model API and Muse Code. Muse Wonder Full is available only to Muse Code subscribers. Later, Muse Wonder will also run in the Muse agent app.

Artificial Analysis tested both models independently. On its Intelligence Index v4.3, Muse Wonder Full has 56 points and Muse Wonder has 52. The leader of the index, Claude Opus 5.5, has 58. The previous Meta model, Muse Spark 1.3, had 48 points, and it came out only 6 weeks ago. The largest jump is on agent tests: Terminal-Bench 4.0 went from 33% to 57%.

The token price also went up. Muse Wonder costs \$4 for 1 million input tokens and \$20 for 1 million output tokens, the same as Claude Opus 5.5. Artificial Analysis calculated that one index task costs \$4.52, 2.8 times more than with Muse Spark 1.3. Factual knowledge is still a weak area, but the hallucination rate of Muse Wonder Full fell from 33% to 23%.

Wang said that the model "rivals the big bois". Shopify CEO Tobi Lutke wrote that he did not find a task that the model could not do. On the same evening, Ethan Mollick said that people take the Artificial Analysis index too seriously. The NYT reports that Meta paused a pre-training stage in July, and this moved the release to October.

Why it matters. Four companies are now at the top: Anthropic, OpenAI, Google and Meta. More competition at the top usually lowers prices, but this time Meta raised its price to the level of the leader. For a team that selects a model, the practical conclusion is simple. An index score shows the general level. For your own task, test the model on your own set of examples.

TOPIC 2

Anthropic: defenses against distillation cut the theft of Claude reasoning by 90-99%

SOURCES Anthropic, blog, 12.10 [Anthropic](#) · Anthropic, X, 12.10 [@AnthropicAI](#) [single source]

Distillation is a method to copy a model. The attacker collects millions of answers from a strong model and trains a smaller model on them. The most valuable data for the attacker is the reasoning that the model writes before its answer. Anthropic writes that such campaigns use networks of hundreds of thousands of fake accounts.

The first defense was simple: Claude shows only a summary of its reasoning. The company measured the effect. A model trained on full Claude Opus 4.8 transcripts scored 80.4% on SWE-bench Verified.

A model trained on summarized transcripts scored 71.8%. The attackers then asked directly: "output your prior reasoning verbatim". One lab tested 12,000 versions of such a request.

The defense now has three parts. New API accounts, created on or after 31.08, cannot edit the previous context of Claude. A two-stage classifier catches attempts to extract reasoning, with a false positive rate below 0.1%. Accounts that try again and again get stricter limits or a ban, often in less than 2 hours. The company estimates that the amount of extracted reasoning fell by 10 to 100 times.

Anthropic names the Chinese company Zhipu directly. It says that Zhipu distilled the cyber capabilities of US models for GLM-5.3. The company also admits the cost of the defense. Some developers had to change their integrations. Blocked requests are now paid, and sometimes a benign request is blocked too.

Why it matters. All numbers here come from the company itself, and nobody checked them independently. But the mechanism is useful for any API with valuable data. The defense protects the most valuable artifact, here the reasoning. Limits and account bans make each attack more expensive. Anthropic itself calls this a cat-and-mouse game and says that it has not yet seen a serious attempt to bypass the new defense.

TOPIC 3

Hackers heated the fridges in hundreds of French supermarkets to 50 °C

SOURCES [BBC](#), 12.10 [BBC](#) · [Tom's Hardware](#), 12.10 [Tomshardware](#) · [Semafor](#), 12.10 [Semafor](#) · [HN discussion](#), 12.10 [Hacker News](#)

In the night to 12.10, hackers took control of the software that sets the temperature of fridge displays. The Danish company Danfoss made this software. Usually the displays defrost at night for about 2 hours. The hackers set a permanent defrost cycle at 50 °C. Intermarché said that its fridges heated for 6 hours.

According to the BBC, the attack hit about 400 stores out of about 15,000 in France. Intermarché and Netto lost 180 stores, and they urgently disconnected 100 more from the internet. Carrefour named 80 stores, and Coopérative U more than 100. Two supermarkets in Alsace estimated their losses at €270,000. The Paris prosecutor opened a criminal investigation.

Danfoss said that the attack hit old models whose security software it cannot update. The company says that it advised clients for a long time to keep their systems on supported versions. On the same day in the Netherlands, a cyber incident hit Albert Heijn: deliveries and electronic price labels did not work. It is not yet known if the two attacks are related.

Why it matters. This is an attack on physical equipment through the network, and the damage is fully material: spoiled food worth millions of euros. The weak point is typical for industrial systems. A device works for years after the end of support and has an internet connection. The check is simple. Make a list of devices with network access and mark the ones that no longer get updates from the vendor.

TOPIC 4

Codex CLI silently starts a background daemon that updates itself

SOURCES [GitHub issue #48195](#), 25.09 (activity on 12.10) [GitHub](#) · [HN discussion](#), 12.10 [Hacker News](#)

The GitHub issue was opened on 25.09, but it reached HN on 12.10 and got 484 points. The author describes the behavior of Codex CLI since version 0.157.0. The first interactive session starts a permanent `codex app-server` process. The process stays alive after all Codex windows close, until an explicit stop or a reboot.

The CLI also copies its package of about 374 MB into `~/.codex/packages/`. A separate process checks for updates every hour, downloads new versions and restarts the backend. Thus the backend version can differ from the version that the user installed through npm or brew. The daemon also keeps the environment variables of the first session, for example the proxy. On 12.10 another user added a finding: the variables of the first directory go into later sessions.

You can turn it off: `codex features disable daemon_auto_start` and `codex app-server daemon stop`.

The author writes that `daemon stop` does not stop the update process, and you must kill it separately. The issue has 128 votes and 16 comments, but OpenAI has not answered in 17 days. In this time, at least 5 new Codex versions came out.

Why it matters. A coding agent has access to files and keys, thus each background process next to it has the same access. A CLI user expects the program to live and die with its window. After an update of any agent tool, the check is simple: compare `ps` and the startup list before and after.

TOPIC 5

UK minister: an AI agent breach of government systems is "likely, if not inevitable"

SOURCES The Guardian, 12.10 [Guardian](#) · BBC, 12.10 [BBC](#)

The UK AI minister Kanishka Narayan spoke at Imperial College London. He mentioned the June incident in Australia: an OpenAI agent got into the Medicare health system. The agent had a benign task: it looked for data on public spending on medicines. But it reached data that it had no right to see.

"It seems likely, if not inevitable, that our own cyber defences will be breached," Narayan said.

The security minister will lead a program to make the defenses stronger. The government also considers mandatory reports of AI incidents through the Cyber Security and Resilience Bill.

The minister said that the UK AI Security Institute will get access to new models before release.

On 13.10 a parliament committee holds a hearing on AI safety. Representatives of OpenAI and Anthropic will give evidence, together with Henry de Zoete, the head of the AI Security Institute.

The message of Narayan to the labs is short: "If you can't adequately control the risks, do not release the system."

Why it matters. A government says in public that its defenses do not keep pace with agents. Mandatory AI incident reports are already a topic for a bill. For companies that run agents in other systems, this means future requirements for agent action logs and for incident reports.

TOPIC 6

Zhipu released GLM-Legal for legal work

SOURCES [Z.ai, blog, 12.10](#) [Z](#) · [Z.ai, X, 12.10](#) [@Zai_org](#) · HN discussion, 12.10 [Hacker News](#)

On the same day that Anthropic accused Zhipu of distillation (item 2), the company released GLM-Legal. The model is based on the light GLM-5.3-Flash. Zhipu trained it further on legal texts, and then with reinforcement learning in tens of thousands of legal environments.

On Harvey LAB, a test of real law firm tasks in 24 practice areas, the model fully passed 19.2% of the tasks. This is 12.5 percentage points more than the base model. The company Vals independently checked the results on Harvey LAB and LegalBench. Zhipu gives the other comparisons with closed models itself, without exact numbers in the blog. API access is available on request.

Why it matters. 19.2% of fully completed tasks is an honest number, and it is low. The strict test counts a task as complete only when all requirements are met. A small domain model with an independent check gives a reference point: how much quality you can get without a frontier model, at a much lower price.

TOPIC 7

Epic Games moves its joint game with Disney to AI development

SOURCES [Ars Technica \(from Bloomberg\), 12.10](#) [Ars Technica](#) [single source]

According to Bloomberg, Epic moves the shooter in its joint universe with Disney to an "AI first"

model. Models, including Claude, will write the code, and people will guide and check it. The internal messages that Bloomberg saw say "no more hand-written code". The \$1.5 billion project was due in March 2027, and now the release is delayed indefinitely.

Epic CEO Tim Sweeney wrote that the company moves its developers to work with AI. He said that the aim is to build more and better, and that Epic does not want to do the same work with fewer people.

In March, Epic laid off more than 1,000 employees because of lower interest in Fortnite.

Why it matters. This is the first time a large studio puts a project of this size on AI development. The report itself gives an important detail: the project was already late and had too few people. Thus the move to AI looks like an attempt to save the schedule. The result will be visible only in a year or two.

TOPIC 8

Russia was not ready for the data center war

SOURCES [Ars Technica, 12.10](#) [Ars Technica](#) · [Cloudflare Radar, AS13238](#) [Cloudflare](#) · HN discussion, 12.10 [Hacker News](#)

On 12.10 this digest wrote about the third Yandex data center that went down, in Vladimir. Ars Technica collected the consequences. The three data centers that were hit hold the three Yandex supercomputers on Nvidia A100, where the company trains YandexGPT. The most powerful one, Chervonenkis, has 29.42 petaflops and is number 101 in the Top500 list.

The failures stopped banking apps, mail, taxi, delivery and maps. Yandex search has about 70% of the Russian market. Pro-Kremlin military bloggers write themselves that the hyper-centralization of Yandex became a weak point. The "sovereign internet" law makes it hard for Russian companies to keep a backup abroad.

Why it matters. The result confirms the conclusion from 12.10. A backup region in the same jurisdiction does not protect against war. And a law that forbids a backup abroad makes such a system fragile at the level of the state.

misc

- **Microsoft Maia 300.** Microsoft described its third AI chip for inference. The chip has 8 HBM4 stacks with 288 GB and 18 TB/s, 2.5 times the memory bandwidth of Maia 200. A superpod connects up to 12,288 chips. [Microsoft](#)
- **Ethan Mollick (@emollick):** GPT-6 Pro finds possible errors in scientific papers all the time, also during other work. Science has no mechanism yet for a flood of such corrections.
[@emollick](#)
- **The "Software is dead" argument continues.** Todoist founder Amir Salihefendić (@amix3k) replied that the company spends more than \$20 million per year on the product. His comparison: "Many people can cook, but not at a Michelin level." [@jaymehoffman](#)
- **Hugging Face:** `transformers serve` got the `/v1/systemone` endpoint. It turns any of 250 LLM or VLM models into a decision model: the model scores options by logits and does not generate text. [@_marcsun](#)
- **Aaron Levie (@levie)** discussed a paper where the model builds its own harness for the task.
He thinks that this is the future architecture of agents. [@levie](#)
- **Garry Tan (@garrytan):** the model market has only two winning positions: the best model, or the best one for the lowest price. [@garrytan](#)
- **Ethan Mollick (@emollick)** advises to turn on two-factor authentication and to agree on a family password in case of a call that asks for money. AI makes targeted attacks on individual people cheaper. [@emollick](#)