

Trump's Super Intelligence Force is led by Jay Clayton

Strata runs a 125B Qwen on a gaming PC, a court voids a sentence over an AI video, a fight over a Claude datacentre.

Monday, 5 October 2026

IN THIS ISSUE

1. "Super Intelligence Force": the White House assembles an AI group that again defers to the companies
2. Strata: a 125-billion-parameter Qwen on a gaming PC with a 12 GB card
3. Arizona: a court throws out a sentence influenced by an AI video of the victim
4. A datacentre for Claude in Queensland: 2.16 GW and 21,500 signatures against
5. RemoveMacAI: how to fully turn off Apple Intelligence on macOS 27
6. AI and jobs: Levie counts new roles, Mollick argues with Cuban

TOPIC 1

"Super Intelligence Force": the White House assembles an AI group that again defers to the companies

SOURCES BBC, 05.10 [BBC](#) · The Guardian, 04.10 [Guardian](#) · Semafor, 04.10 [Semafor](#)

On Sunday Donald Trump announced that Director of National Intelligence Jay Clayton becomes the new White House AI czar and will lead a group the president called the "Super Intelligence Force".

According to the BBC, its members include Federal Trade Commission Chair Andrew Ferguson, Undersecretary of Defense for Research and Engineering Emil Michael and Office of Personnel Management Director Scott Kapor. Semafor adds the vice president, the defense and Treasury secretaries and David Sacks, who was AI czar until March. The group reports directly to the president and to Chief of Staff Susie Wiles. Its job, per Trump's post, is to coordinate the government's engagement with consumers, public interest groups, religious organizations, critical infrastructure providers and "super intelligence companies".

Yesterday this digest noted that WSJ gave the group 120 days to report, while Clayton's appointment was still only a rumour. Now it is official, and Semafor spells out the content: the report has to assess the risks and opportunities of AI and define the federal government's responsibility. Clayton told WSJ he expects "mechanisms developed in dialogue" with companies, but according to Semafor, citing other officials, the group will

leave risk management to the firms themselves. The Treasury secretary repeated the main point on Saturday: "We cannot lose our lead to the Chinese."

The reactions are telling. Senator Mark Warner, vice-chair of the intelligence committee, cautiously welcomed the appointment, but noted that the administration "has resisted meaningful safeguards without offering a serious alternative". Elon Musk announced the same day that he is renaming SpaceX's AI unit from SpaceXAI to SpaceXSI. And Sam Altman, in an interview with Politico published on 04.10, said "the world should accept some bad things happening for the benefits of this technology". The administration fixed the term "super intelligence" in place of AI with a separate executive order; this digest covered the rename on 23.09.

Why it matters. The group is made of intelligence, defense, finance and the competition regulator, so in Washington AI is now first of all a national security matter. For companies that sell in the US the practical signal is that there will be no hard rules in the next 120 days, and the report will be shaped by the same people they already agreed a "morally binding" pact with on 29.09.

TOPIC 2

Strata: a 125-billion-parameter Qwen on a gaming PC with a 12 GB card

SOURCES Strata repository on GitHub [GitHub](#) · Qwen3.8-Flash-Next model card on Hugging Face [Hugging Face](#) · HN discussion, 04.10 [Hacker News](#)

Strata is a free MIT-licensed installer and engine that runs Qwen3.8-Flash-Next on an ordinary computer. The model itself came out in August: 125 billion parameters, of which 6 billion work on each token, plus 51 billion parameters of n-gram embeddings. Inside there are 512 experts per layer, of which 10 routed and one shared expert switch on for each token. Strata exploits exactly this architecture: it keeps the most frequently used experts in video memory, all the others in system RAM, the processor computes the rest in parallel, and the large embedding table sits on the SSD.

Numbers from the author's README: on an RTX 5070 with 12 GB and 64 GB of RAM the most compressed version (Q2_0) writes 94 tokens per second, and the highest-quality one available (IQ3_S) 53. On an AMD RX 9070 XT it is 60 tokens per second. A commenter on HN with an RTX 4090 and 128 GB of memory reported 124 tokens per second.

Requirements: 12 GB of video memory or more, 32 GB of RAM or more and about 80 GB of disk; the model download is about 70 GB. There is a separate Coder version with half the experts removed: according to its authors it reaches 91% of the full model's SWE-bench Verified score and fits in 32 GB.

The repository was created on 24.09 and in 11 days collected more than 11,000 stars and 845 commits, many of them written together with Claude. The skepticism on HN is about quality: the models are compressed to 2-3 bits per parameter, and nobody has published

how much intelligence they lose in the process. Another remark: the README suggests pasting "set up Strata following the instructions in the repository" into a coding agent, and one commenter compared it to `curl | bash`, only worse.

Why it matters. A year ago a model of this size needed server GPUs. Now running it locally at tens of tokens per second costs as much as a gaming PC, and for tasks where data cannot go to the cloud this is a real option. Before relying on it, the compressed version is worth testing on your own tasks: the generation speed has been measured, the quality at 2 bits has not.

TOPIC 3

Arizona: a court throws out a sentence influenced by an AI video of the victim

SOURCES The New York Times, 04.10 [NYT](#) [single source]

An Arizona appeals court upheld the manslaughter conviction of Gabriel Horcasitas but threw out his sentence of 10.5 years in prison. In 2021, in a Phoenix suburb, he shot 37-year-old Christopher Pelkey dead during a road-rage incident. At sentencing, the victim's sister played a video generated with AI in which Pelkey tells his killer "I believe in forgiveness". The voice was taken from a YouTube clip, the face from a funeral poster, and the script was written by the sister herself. The judge said "I loved that A.I.", called the video genuine and imposed the maximum sentence, more than the nine years prosecutors had asked for.

The three-judge panel called this a "fundamental error". In their words, the video does not reflect actual events but presents its statements as if they came directly from the victim, and it "erases the interpretive distance between the family's belief about what the victim would have said and the victim's own voice and opinions". A victim's right to be heard, the judges write, cannot trample a defendant's right to be sentenced on accurate and reliable information. Horcasitas will now be resentenced.

Why it matters. This is one of the first cases where a higher court has directly weighed a synthetic video in a courtroom, and the ruling is about influence: the video was openly labelled as AI and still affected the sentence. For lawyers and judges the argument is practical. A reconstruction of a person's voice already carries the weight of testimony, even though it is not testimony.

TOPIC 4

A datacentre for Claude in Queensland: 2.16 GW and 21,500 signatures against

SOURCES The Guardian, 04.10 [Guardian](#) [single source]

Near Dalby, 180 km west of Brisbane, Australia's largest datacentre is planned, worth \$31bn and built for Anthropic's models. The developer is Singapore-based Zerra DC, and the 725-

hectare site is currently a 24,000-head cattle feedlot surrounded by gas wells. Peak capacity is 2.16 GW, about a quarter of the whole state's energy demand. The developer promises a cooling system that uses about as much water as an average shopping centre, and around 1,400 long-term jobs.

Locals remember the boom and bust of coal seam gas in the same region and see the same story. A petition against the project collected more than 21,500 signatures in a few weeks. The Environmental Defenders Office says queries about datacentres to its legal advice line are up 250% since January and calls for a moratorium. On Friday the state government took decisions on datacentres away from local councils and moved them to the state level, although the mayor of Dalby had expected only a month ago that the project would stay with the council.

Yesterday this digest noted that on Tuesday executives from OpenAI, Anthropic, Microsoft and Google will appear before an Australian parliamentary committee on AI. The Dalby datacentre will most likely come up there too.

Why it matters. Resistance to datacentres is no longer only an American story: according to the Guardian, three quarters of Americans oppose them and more than 500 counties have imposed bans. The bigger the project, the more the decision moves from the local level to the state, and the longer the construction timelines the labs are counting on become.

TOPIC 5

RemoveMacAI: how to fully turn off Apple Intelligence on macOS 27

SOURCES RemoveMacAI repository on GitHub [GitHub](#) · HN discussion, 04.10 [Hacker News](#)

According to the utility's author, macOS 27 no longer has a single switch for Apple Intelligence, and the models stay on disk after the features are turned off. RemoveMacAI does three things: it turns the features off (Siri, Writing Tools, Genmoji, Image Playground, the ChatGPT extension, summaries in Mail, Messages, Safari and notifications, text predictions, code completion in Xcode), removes the models through Apple's own asset service and stops the system from downloading them again. For that last part, a configuration profile redirects the download of each model to a closed local port.

System Integrity Protection stays enabled, and everything can be reverted with one command.

The thread collected 402 points and 265 comments on HN. The author also explains details that confuse users: Storage settings keep counting Apple Intelligence for a while after removal because the system deletes the files on its own schedule, and a process named "Siri" keeps running because in macOS 27 that is the name of the Spotlight window.

Why it matters. AI built into the operating system now takes disk space and is harder to switch off than on, and some users want control over it. For companies with managed Macs something else matters more: the same can be done with Apple's standard restriction keys in a profile, without third-party scripts.

TOPIC 6

AI and jobs: Levie counts new roles, Mollick argues with Cuban

SOURCES @levie on X, 05.10 @levie · @emollick on X, 05.10 @emollick [single source for the numbers]

Box CEO Aaron Levie quoted a LinkedIn estimate relayed by The Kobeissi Letter: since 2023, more than 750,000 AI-linked jobs have appeared in the US, and the largest group among them, 282,000, are data annotators. Levie adds that the statistics undercount people inside large companies who were simply moved to AI work: banks, life sciences companies, manufacturers and even law firms are hiring engineers to deploy agents.

Mark Cuban replied with the classic "AI won't take your job. Someone who knows how to use AI better than you will take your job." Ethan Mollick disagreed on three points. It is unclear which skill of "using AI better" stays valuable as AI gets easier to use. More and more often it is organizational systems that bottleneck AI gains, and the individual worker is no longer the main factor. And the effect on jobs is uncertain and may hit entire job categories. Mollick's conclusion: work systems should be built so that AI augments people, with automation as only one of the options.

Why it matters. Both sides describe something real, and the LinkedIn figure confirms it: most of the new jobs appeared in serving models, data annotation and datacentres. For someone planning a career, watching how processes change inside their own company is more useful than simply learning to "prompt better".

misc

- **OpenAI's legal risks are growing.** Semafor and the FT write about the legal fallout from incidents involving the company's agents: more than 100 organizations have already been notified.

A continuation of yesterday's story about the review of 50 PB of records. The FT piece is paywalled. [Semafor](#)

- **WSJ: AI spending is becoming almost impossible to budget.** According to a study the paper cites, only 11% of nearly 400 businesses were able to forecast their AI spending accurately. A follow-up to yesterday's item on Willison's spending caps. Only the first paragraph is available [single source].

[WSJ](#)

- **The Tarbell Center is committing \$10 million to AI journalism** by the end of 2027. The first grant, \$480,000, went to Machine Gods, NPR's new video podcast from former Hard Fork hosts Kevin Roose and Casey Newton. [Semafor](#)
- **@lennysan** released a conversation with Tibo Sottiaux, who leads ChatGPT and Codex at OpenAI:

why the main bet is now the Dots personal agent and why loops and graphs in agent workflows are, in his view, a passing phase. The transcript is paid. [@lennysan](#)

- **@emollick** got a browser-based brutalist city builder out of Fable 5.1 from a single prompt and compared it with what GPT-5 could do a year ago. The code is on GitHub. [@emollick](#)
- **@emollick** on agent security: he wants an operating system built for working alongside AI, with fine-grained permissions, visibility into what the agent is doing and separate auditor agents. A reply to Sriram Krishnan's post that giving an agent full disk access is a sign of an "awkward intermediate era". [@emollick](#)
- **@garrytan** on everyone building the same thing (agent loop, memory, sandboxes, connectors): it is a sign of real demand, and the market will eventually converge on one "correct OS". [@garrytan](#)
- **A botched redaction revealed water and electricity use at Google's datacentre in Lincoln, Nebraska:** 52.65 MW of peak demand and 13.3 million gallons of water a year. The text under the black box could be copied with an ordinary selection. The story is from 30.09; it took off on HN yesterday. [1011now](#)
- **Tyler Cowen wrote a blurb for Mollick's new book "Coexistence"** and addressed it to the models themselves. [Marginal Revolution](#)
- **headstart** starts compiling dependent Rust crates before their dependencies finish type-checking, and according to its author makes builds up to twice as fast. [GitHub](#)