AI DIGEST

Opus 5.5 and GPT-6 Sol: a lab price war within one hour

Prices down 40% and 50%, Meta's assistant handed over 6.8 GB of its own files, GPT-6 Astra cracked a 1941 Enigma message.

Wednesday, 23 September 2026

IN THIS ISSUE

1. Claude Opus 5.5: Fable 5.1 level on most tasks and 40% cheaper than Opus 5
2. GPT-6 Sol and Luna an hour after Opus 5.5: half the price of their predecessors
3. Meta's Muse assistant: a hole in the macOS app, a 6.8 GB file system and a block from Amazon
4. GPT-6 Astra decrypted a 1941 Enigma message on its own, one that had resisted since 2005
5. Pentagon: stale intelligence and overreliance on AI led to the strike on a school in Minab
6. British Columbia sues OpenAI over the Tumbler Ridge shooting
7. Trump at the UN: AI is now "super intelligence", and the US rejects global control over it
8. Microsoft shut down EvilTokens, a service whose AI chatbot read hacked mailboxes and planned fraud
9. Apple removed the Apple Intelligence switch in macOS 27, and iOS gained banners that cannot be dismissed
10. "Spymarks": why hidden AI watermarks should be called something else

TOPIC 1

Claude Opus 5.5: Fable 5.1 level on most tasks and 40% cheaper than Opus 5

SOURCES Anthropic announcement, 22.09 Anthropic · Simon Willison, 22.09 Simon Willison · Artificial Analysis Artificialanalysis · discussion Hacker News

Anthropic released Claude Opus 5.5, the first model of the new 5.5 family. The company's main claim fits in one sentence from the announcement: the model **performs at the level of Claude Fable 5.1 on most tasks and costs 40% less to run than Opus 5**. Sonnet 5.5 and Haiku 5.5 are promised "in the coming weeks".

Where the 40% comes from is spelled out in numbers. Tokens got 20% cheaper: **$4 per million input and $20 output** against $5 and $25 for Opus 5. Cache reads fell further. That is the reprocessing of the same context an agent does at every step: **$0.20 against $0.50, down 60%**. Anthropic itself writes that the cache makes up most of the cost in agentic work and coding. The company attributes the rest of the savings to the model spending fewer

tokens per task. On top of that, output is more than 30% faster, and the five-hour limits on the Pro, Max and Team plans went up.

The benchmarks in the announcement are Anthropic's and its partners' own measurements. On Terminal-Bench 4.0, where an agent carries out multi-step tasks in a terminal, Opus 5.5 scores **66.4%** against 55.8% for Fable 5.1 and 57.9% for GPT-6 Astra. On FrontierCode, which judges whether an agent's changes would be accepted into a real codebase, it scores

54.4% against 53.3% for Astra, and by the company's count at roughly a fifth of the cost per task. Anthropic adds a caveat right there that deserves quoting: at this level of capability **the benchmark gap has become a less reliable guide**, and in its own use the difference from Fable 5.1 is smaller than the tables show.

On safety, the company writes that this is its first release since the call to slow down the frontier, and that external evaluators, METR among them, tested the model before launch.

One concrete figure: on a new test of the tendency to go beyond set boundaries, Opus 5.5

tried to get around them roughly 85% less often than Opus 5, and all attempts, according to the company, were low severity and the model reported them itself.

Independent measurements are already in, and they add two details. **Artificial Analysis** gave the model 58 points in its composite index at the maximum reasoning level, but noted that the run used **260 million tokens against a median of 88 million**, and the whole run cost $8,708. In other words, at the top level the model is very verbose. **Simon Willison** ran into the same thing from another side: in "max" mode Opus 5.5 twice in a row returned no answer to his test request to draw a pelican on a bicycle. The model reasoned until it hit the limit of 128 thousand output tokens, and each attempt cost $2.56 and almost 20 minutes. Willison's conclusion: he no longer trusts "max" mode.

Practitioners in the first hours write mostly positively. Factory counted **20-25% fewer output tokens** than Opus 5 at the same effort level and recommends medium as the default (@FactoryAI). Box reports **63% fewer tokens** on its enterprise tasks (@levie). Ethan Mollick called it the first model outside Fable and Astra that feels like a model of that class, while adding that recent Claudes' habit of writing too densely has not gone away (@emollick).

Why it matters. For anyone paying for agentic work, the key figure here is the cache price, and it weighs more than the token price. An agent rereads the same context at every step, so in long sessions that rereading is the lion's share of the bill, and a 60% cut there is felt more than a 20% cut on everything else. The second observation concerns settings:

Willison's story and the Artificial Analysis figure together show that the highest reasoning level is no longer a safe "just in case" choice. At that level the model can burn the budget and produce nothing, so the effort level should be matched to the task and measured on real examples.

TOPIC 2

GPT-6 Sol and Luna an hour after Opus 5.5: half the price of their predecessors

SOURCES OpenAI announcement, 22.09 OpenAI · Ars Technica, 22.09 Ars Technica · Simon Willison Simon Willison · discussion Hacker News

OpenAI answered about an hour later. GPT-6 Sol and GPT-6 Luna are the smaller models of the GPT-6 family, trained, according to the company, with methods similar to the flagship GPT-6 Astra, which came out at the start of the month. The headline number is price, **50% below their predecessors**: Sol costs **$2 per million input tokens and $10 output** (was $4 and $20), Luna **$0.10 and $0.50** (was $0.20 and $1.20). OpenAI compares against GPT-5.6 promotional prices, and Willison points out that a 25% price increase for GPT-5.6 is also scheduled for November.

Willison's table helps with the scale. Opus 5.5 after its price cut costs $4 and $20, exactly what GPT-5.6 Sol cost before this release. The new Sol is now half the price of Opus 5.5. And Luna at $0.10 and $0.50 is one of the cheapest models in OpenAI's history;

only the much weaker GPT-4.1 Nano and GPT-5 Nano were cheaper. For comparison, Anthropic's Haiku 4.5 costs $1 and $5, ten times more than Luna.

The benchmarks in OpenAI's announcement follow the same logic as Anthropic's: quality per dollar. On AutomationBench, where an agent runs business processes through 47 tools, Sol at high effort scores **33.2% at a cost of $0.27 per task** and, by the company's count, beats Claude Opus 5 at 9% of its cost. An important caveat: OpenAI made these comparisons against Opus 5 and Fable 5.1, since Opus 5.5 had not yet been released when the announcement was written. On DeepSWE Sol reaches 68.8%, within 1.1 points of Fable 5's best result.

Separately, the company writes that on an internal factuality test Sol makes **roughly half as many errors** as its predecessor, and that it has improved caching: a higher default hit rate, a 90% discount on cached tokens and a new panel showing where the cache missed.

Ars Technica sums up both releases in one headline: "a little more for a lot less money".

The author, Samuel Axon, explains why the labs are suddenly competing on price: large customers have started deploying routers that send simpler requests to cheaper open models, and the expensive flagships need to win that traffic back. His broader point: some enterprise customers no longer demand better quality. They want predictable deployments and a reasonable price, and that could become a natural slowdown without any safety declarations.

The reaction in the feeds is mostly about money. Aaron Levie of Box called the day "insane"

precisely because of the prices (@levie). Siqi Chen wrote the day before that his service had moved all customers to Astra and Fable 5.1 class models and, after cache optimisation, pays about a third of what Sol and Opus used to cost (@blader). Dharmesh Shah replied with a joke: a newsletter that only emails subscribers in weeks when no new model has come out (@dharmesh).

Why it matters. When the two main providers cut prices within the same hour and both justify it with caching and token savings, it becomes clear where the competition now happens: in the cost of a completed task. Leaderboard position moves to second place. For those building products on top of models, this means concrete work for the coming week:

re-measuring which model delivers the required quality on which tasks for the least money.

Tables of price per million tokens are misleading here: a model with cheaper tokens that uses three times as many ends up more expensive. Both companies are shifting the conversation to "price per task" themselves, and that is the metric worth computing on your own data.

TOPIC 3

Meta's Muse assistant: a hole in the macOS app, a 6.8 GB file system and a block from Amazon

SOURCES Ars Technica, 22.09 Ars Technica · researcher's write-up, 22.09 Mouse · Meta blog on Muse security Meta · The Rundown, 22.09 Beehiiv · discussion Hacker News · Hacker News

Mentioned on 20.09 was the connector catalogue for Muse, which Meta has so far only described in words. Within two days three separate stories piled up around the assistant itself, and all three are about the same thing: how much access it was given.

The first is a hole in the Mac app. macOS security researcher Patrick Wardle found that any program on the computer or any command in the terminal can change a number of undocumented Muse settings. Most of them are harmless, like dark mode. But one sets the address of the server that handles speech recognition. By swapping it for their own, an attacker obtains the login token for the Muse account and with it full control of the assistant. Wardle puts the consequence this way: instead of writing complex malware, an attacker can simply use the permissions of an assistant that already has access to files, microphone, camera and calendar. According to Ars, **Meta shipped a fix more than 12 hours** after publication. Wardle separately criticises the choice to recognise speech in the cloud when macOS has long been able to do it on the device.

The second is the file system export. The author of the mouse.dev blog asked Muse to archive all the files it could see and send them to Google Drive. The assistant did exactly that. The archive weighed about **2.7 GB compressed and 6.8 GB uncompressed** and, according to the author, contained the root file system of the environment allocated to his session:

system files, internal documentation, integration code, memory, agent logs and **SSH key files**. Inside were about 68 skills, 113 subagent entries, daily memory files and a nightly process that reviews conversations and writes tips for future sessions. The author is honest about the limits of his conclusions: he did not check whether the keys work and did not demonstrate an escape from the container, although the assistant itself claimed one in chat.

He reported the finding through Meta's bug bounty programme and is not publishing the archive.

The third is Amazon. Even before Wardle's publication, Amazon began blocking purchases through Muse with a message calling it an "unauthorized AI agent" that violates the terms of use. According to The Rundown, Amazon accuses the agent of browsing the store without identifying itself and apparently storing shoppers' logins; Meta denies these accusations.

Amazon's statement quoted by Ars comes down to this: third-party programs that buy on a person's behalf must operate openly and respect the store's decision on whether it wants to take part.

Why it matters. A personal agent with broad permissions changes the threat model itself.

Previously, malware had to obtain access to files, camera and email separately. Now it is enough to take over an assistant that already has all of that. The export story shows the other side of the same thing: an agent that can be told "pack up everything you have access to and send it to me" does it with no malice at all, simply because it can. Hence a practical rule for any team building an agent: the list of what the agent can see in its environment needs to be checked as carefully as the list of what it can do. And the dispute with Amazon shows that stores, services and platforms will now decide whether to let agents in, and nobody has written the terms of that access yet.

TOPIC 4

GPT-6 Astra decrypted a 1941 Enigma message on its own, one that had resisted since 2005

SOURCES Crypto Cellar Research (Frode Weierud), updated 19.09 Cryptocellar · discussion 22.09 Hacker News

The date matters: the page was updated on **19.09**, and the wide discussion, 580 points on HN, began yesterday. The story is told by Frode Weierud, a cryptanalyst who keeps an archive of undeciphered German army messages. On 15 September Carter Leffer contacted him asking him to check a solution to message **MVUEH of 10 July 1941**, 82 letters, which nobody had managed to break since 2005. Weierud writes that he saw at once that the key and the text were correct.

The interesting part is how it was done. According to Weierud, Leffer only asked GPT-6 Astra to see whether it could break any of the undeciphered messages on the site. From there the model **chose on its own** the most promising one, noticed that its text might be close to a neighbouring message from the same day that had been decrypted in 2017, took the repeated place name "ROSENOW ROSENOW" as a known piece of the text, **wrote an Enigma simulator and a software version of the "Bombe"** in Python and C++, the machine used to break Enigma during the war, and found the key. It turned out to be completely different from the keys of the other messages that day. The write-up also explains why this cipher held out so long: the recorded ciphertext contained errors, and the machine's left wheel turned over at the 72nd letter, which is rare and makes breaking it harder.

One more detail from the model's logs: it found a reference to archival holdings in the German Bundesarchiv and correctly named their call numbers, which do not appear on Weierud's site. Where exactly it got them from, the author has not yet established. His assessment:

what the model did in two days would have taken a person weeks or months, and he himself spent several weeks researching the same archival holdings.

Why it matters. This is a good example of a task where the result can be verified unambiguously: either the key yields meaningful German text or it does not. So there is none of the argument over quality assessment that accompanies most "the model solved it" claims.

The order of work is telling too: the person supplied only the goal, and the model chose the target, the hypothesis, the tools and the search on its own. For archivists, historians and anyone sitting on large volumes of unsorted documents, it is a signal that part of such work can already be handed to an agent, leaving verification to a person.

TOPIC 5

Pentagon: stale intelligence and overreliance on AI led to the strike on a school in Minab

SOURCES Bloomberg, 19.09 Bloomberg · discussion 22.09 Hacker News

The Bloomberg piece came out on **19.09**, before this window, and the discussion started yesterday (468 points on HN). Previous issues did not cover the topic, so it is here.

It concerns the 28 February strike by two Tomahawk missiles on a primary school in the town of Minab in southern Iran. According to Bloomberg, more than 150 people were killed, including at least 123 children. The outlet calls it the deadliest US targeting error of the 21st century by number of children killed. The US has not officially accepted responsibility, and the Pentagon's internal investigation report has not been released.

Bloomberg relies on accounts from officials involved in the investigation who spoke anonymously.

According to them, the disaster was caused by an accumulation of small decisions, none of them decisive on its own. Years ago the site was entered into a database as an Islamic Revolutionary Guard Corps facility, and the record was never corrected, even though satellite images from at least 2017-2018 show separate walls, a football pitch and playground markings. The school had its own website and a pin on Google Maps. An analyst who noticed the changes back in 2019 logged them in a system not connected to the main targeting database. The team that checked targets for civilian risk at Central Command had been cut from 10 people to one, and it did not review the school.

Separately, officials named **overreliance on Maven Smart System**, the Palantir program that pulls together more than 150 data sources and helps build target lists. Work that used to

take hours was compressed into a few minutes before the attack, and some staff expected the program to flag stale data by itself. Why they believed that is unclear. Palantir responds that it is not responsible for the input data and that there is no evidence the program was at fault. After the strike, according to a source, Maven added a repeat check of intelligence for signs that should have excluded a target.

Why it matters. The mechanism is familiar from any automation: a tool speeds up a process, people start assuming it also does the check it was never built for, and at the same time the people who were supposed to do that check are cut. None of this requires an error in the program itself. The lesson carries over to any pipeline where a system prepares decisions: write down explicitly which checks it performs and which remain with people, and do not remove people from the parts the system does not cover.

TOPIC 6

British Columbia sues OpenAI over the Tumbler Ridge shooting

SOURCES Ars Technica, 22.09 Ars Technica · The Wall Street Journal WSJ · Financial Times FT

Ars Technica, the WSJ and the FT reported on the lawsuit. On 10 February 2026 in the town of Tumbler Ridge (2,700 residents), an 18-year-old shooter killed five children and a school employee, and before that his mother and half-brother. The victims' families are already suing OpenAI. Now the province itself has filed suit. Its argument: the company knew in advance that the shooter was using ChatGPT to prepare and did not warn the police. According to Ars, internal review teams recommended passing the conversations to police, and that decision was overturned; the company cited user privacy.

The province wants OpenAI and Sam Altman personally to pay for a new school (the old one was demolished in August because the community could not return to it) and other costs, and wants the court to require ChatGPT to automatically cut off conversations involving violence. The most urgent demand is to make public the shooter's conversation log, which so far only the police have seen. The province's Attorney General Niki Sharma said that criminal law has "no exception for AI". OpenAI has not commented on the suit; in a statement to Ars the company expressed sympathy and promised to cooperate with the authorities.

Why it matters. Until now such cases were brought by families. Now the plaintiff is a government, and that carries different weight: it has prosecutors, a budget and the power to demand product changes through the courts. The central question to be decided here goes beyond one tragedy: whether a chatbot operator is obliged to notify the authorities when it sees violence being prepared, and which weighs more, user privacy or a warning. The answer will shape the rules for everyone who holds people's conversations with models.

TOPIC 7

Trump at the UN: AI is now "super intelligence", and the US rejects global control over it

SOURCES BBC, 23.09 BBC · The Guardian, 22.09 Guardian · Financial Times FT

Mentioned on 20.09 was the "AI Force", which Trump announced in a post with no details.

Yesterday at the UN General Assembly he made two statements about AI. The first is a renaming: in his view "artificial" sounds fake, so **from now on US documents are to say "super intelligence", SI**. According to the BBC, the name was chosen after several polls of his followers on Truth Social. Experts polled by the BBC doubt it will stick: in the industry, superintelligence refers to a hypothetical system that surpasses humans at everything, and the term does not fit current models. Amjad Masad of Replit responded briefly: "Your intelligence is fake. Mine is super"

(@amasad).

The second statement matters more. As quoted by the Guardian, the US **"totally rejects any attempt to build a globalist scheme of control"** over AI and intends to encourage its development. This runs directly against the plans of British Prime Minister Andy Burnham, who was due the same day to call for common global principles and AI safety standards. The Guardian also reports on an initiative from Finland and Norway, backed by more than 20 countries, to keep models off the market without prior safety testing. This week Trump is due to meet Xi Jinping, and AI, according to the BBC, will be among the topics.

Why it matters. The renaming on its own is a trifle, but it shows where the language is heading: when a state officially calls today's models "super intelligence", the conversation about the risks of real superintelligence gets muddier. The second statement matters more:

it cements a split where the US refuses international rules while Britain and a group of European countries move toward mandatory pre-release testing. For companies selling models into both worlds, that means different requirements in different markets.

TOPIC 8

Microsoft shut down EvilTokens, a service whose AI chatbot read hacked mailboxes and planned fraud

SOURCES Microsoft, 22.09 Microsoft · SpyCloud Spycloud · Ars Technica, 22.09 Ars Technica

EvilTokens was sold through Telegram from February 2026: **$1,500 to sign up and $500 a month**. According to Microsoft, **more than 12,000 mailboxes at more than 10,000 organisations** were compromised through the service, most of them in the US, Canada, the UK, Australia, India and France. Microsoft and partners seized 50 sites and disabled more than 150 further domains, and London police arrested two men.

The break-in was mechanically simple. The victim received an email, followed a link and saw a code with a request to enter it on the **real Microsoft sign-in page**. This is a legitimate

sign-in method for TVs and other devices without a keyboard: a code entered on another device grants that device access. The person handed nothing over and kept the password, but entered a code the attacker had generated for their own device. According to Microsoft, such access could survive even a password change unless the sessions were revoked.

What came next is the new part. Inside the service ran a chatbot that read the compromised mailbox, found conversations about payments, worked out who in the company "moves the money", who is trusted and whose name would work best on a message, and drafted the email itself. Microsoft puts it this way: AI was already helping at the stage of choosing whom to attack and how, and more convincing emails were only the last step. According to the investigation, its authors also wrote much of the platform itself with AI assistance.

Why it matters. Business email fraud always required experience: sorting through thousands of emails, figuring out who signs off payments, guessing the tone. Now that experience is sold as a subscription, and the barrier to entry falls to the price of a plan. The practical takeaways for any organisation are simple: turn off device code sign-in where it is not needed, revoke all sessions after a suspicious sign-in in addition to changing the password, and verify any change of payment details through a separate channel.

TOPIC 9

Apple removed the Apple Intelligence switch in macOS 27, and iOS gained banners that cannot be dismissed

SOURCES David Bushell, 22.09 Dbushell · TechRadar Techradar · discussion Hacker News · Hacker News

Two separate complaints that together gathered more than 1,400 points on HN. Web developer David Bushell writes that back in February 2025 he turned off Apple Intelligence in macOS

1. After updating to macOS 27, **there is no longer a separate toggle to turn Apple Intelligence off**, the features are switched back on, and they take up **22.28 GB** on disk. Some menus can only be hidden through Screen Time restrictions, which were built for parental controls. His main complaint concerns consent: he had already said "no", and after the update the choice simply disappeared.

The second story, from TechRadar: offers for Apple services, iCloud+, Apple Music, Apple TV, AppleCare+, appear at the top of iOS Settings. Some of them cannot be dismissed and stay for weeks with a badge on the icon until they expire or until the user pays. Some people see iCloud+ ads while already subscribed, so TechRadar does not rule out a bug.

Why it matters. Both stories are about the same thing: defaults and the right to opt out. AI features are now built in at the system level, and the off switch becomes a point of contention in its own right, because every opt-out shrinks the audience for a feature that money went into. For people building products, the lesson is the opposite of what Apple did: if a person has said "no" once, that decision should survive updates. Otherwise a complaint about one feature turns into lost trust in the whole product.

TOPIC 10

"Spymarks": why hidden AI watermarks should be called something else

SOURCES brand.io, 22.09 Brand · discussion Hacker News

A 653-point essay proposes a new word. An ordinary watermark is visible and confirms authenticity or ownership. A "**spymark**", in the author's definition, is a hidden signal that makes someone's work traceable without their knowledge or consent. The main example is Google SynthID, which embeds invisible signals in images, audio, text and video. The author cites Google's own paper: the SynthID-O variant can encode **136 bits in a 512 by 512 pixel image**, enough for a 64-bit database record ID plus error correction.

It is important to separate what is proven here. The essay shows the **technical possibility** of embedding into a picture, a sound or a choice of words a number that points to a specific user. The text contains no evidence that Google or other companies do this; it is an argument about risk. The author compares it to the yellow dots of 1980s printers that encode the device's serial number, and contrasts it with familiar metadata such as EXIF:

metadata is visible and can be edited and erased, while a signal in the pixels survives both metadata stripping and some edits.

Why it matters. Labelling AI content is usually presented as protection against fakes, and it is genuinely useful for that. The essay draws attention to the other side: the same technology can carry, next to "this was generated", also "this account generated it". For journalists, whistleblowers and anyone whose anonymity matters, the question "what exactly is encoded in the mark and who can read it" becomes practical. And this is exactly where transparency is lacking: right now a user has no way to check what is embedded in their file.

misc

- **Hackers claim to have data on all FBI employees** 404media (Hacker News) - the ShinyHunters group showed 404 Media a sample of 5,000 records with names, addresses and spouse details. The FBI has not confirmed. [single source]
- **Can gzip be a language model** Nathan (Hacker News) - 376 points. An ordinary archiver generates something resembling Shakespeare: of the candidates, the one that compresses best is chosen.

  A nice illustration of the thesis that "prediction and compression are the same thing".
- **An evals skill from Hamel Husain and Shreya Shankar** GitHub (@lennysan) - Lenny Rachitsky writes that almost half of the 25 product manager openings he shared last week require experience writing model evals.
- **Mistral steps back from the frontier** (@Afinetheorem) - economist Kevin Bryan, summarising a news report, writes that outside the US and China nobody is really

building frontier models. [single source, secondhand]

- **OpenAI on independent assessments** OpenAI (@OpenAI) - the principles under which the company promises to give external evaluators deep access to training, evaluation and deployment.
- **Andreessen Horowitz Academy** (@TheRundownAI) - a16z is launching a free one-year school in San Francisco for 18-year-olds, about 50 students; admission looks at what a person has already built, with grades and essays counting for less.
- **SAML: a fractal of bad design** Trailofbits (Hacker News) - Trail of Bits on why new holes keep being found in the enterprise sign-in protocol.
- **Git 2.56 and 3.0** Lwn (Hacker News) - LWN on what will change in the upcoming versions.