AI DIGEST

ChatGPT carries an ad identifier onto other sites

The mechanics across 936 pixels, US-China AI talks, $300bn of exposure off balance sheets.

Monday, 21 September 2026

IN THIS ISSUE

TOPIC 1

ChatGPT carries an advertising identifier onto third-party sites - the mechanics

SOURCES research Buchodi · discussion Hacker News

An independent researcher reverse-engineered OpenAI's ad collector at `bzr.openai.com` step by step (`bzr` stands for `bazaar`, OpenAI's internal name for the ads platform). The mechanism was reproduced on the author's own phone, confirmed with **two independent capture methods**, and cross-checked against several months of traffic: **936 advertiser pixels across 1029 hostnames**.

How it works. On `chatgpt.com` the client generates 16 random bytes and asks the backend to sign them. Back comes an RS256-signed JWT that **binds the advertising identifier `obi` to the `sub` field, meaning the user's account**. The token lives for 60 seconds. The identifier then settles as a `__obi` cookie on the `.openai.com` domain with a one-year `Max-Age` and the flags `SameSite=none; Secure` - exactly the combination that lets cookies ride along on cross-site requests.

This is where it gets interesting. Any advertiser installs a small piece of OpenAI code on its own site, the same way retailers have installed the Meta or Google pixel for years. And `__obi` travels to OpenAI together with data about the page: what a person is searching for, reading, buying.

One detail deserves a separate stop, because it shows the limit of "private" modes.

The SDK has a code path that deliberately omits credentials. **It does not help**: the browser attaches cookies to the `<script src>` request itself, that is BEFORE a single line of OpenAI's code runs. Merely loading the tag already discloses the identifier.

And one more number that reframes the conversation. The SDK collects identity from the advertiser's page from four sources, labelled by OpenAI itself: `in` for what the advertiser passes deliberately, and `fm`, `ht`, `js` for what the SDK **scrapes on its own** from form fields, rendered page text, and the tag-manager bus. In observed traffic, scraped identity outnumbered deliberately supplied identity **685 events to 255**. The largest source of email addresses is the tag-manager bus, whose push function the SDK replaces with its own.

Why this matters. The mechanism itself is ordinary adtech, about fifteen years old.

What is unprecedented is where it now runs: on a product people tell things they would never type into a search box. Into a search field you enter a query; into a chat you narrate a situation, with diagnoses, debts, conflicts and documents. Binding `obi` to `sub` means this is the same account, not an anonymous profile. The practical takeaway for any product with a chat in it: the privacy boundary is defined by which cookies actually travel to third-party hosts, and that is verified by capturing traffic rather than by reading a policy.

TOPIC 2

The US and China agree to talk about AI ahead of the Trump-Xi meeting

SOURCES FT FT · NYT NYT

Treasury Secretary Scott Bessent met his Chinese counterpart He Lifeng in New York.

The two sides agreed on a separate dialogue track for artificial intelligence, ahead of Xi Jinping's state visit to the United States.

Yesterday we mentioned Trump's post about an "AI Force" with a promise to appoint an "AI czar", carrying no detail at all. Today it is visible what sits behind the rhetoric: while the public part looks like a social media post, the working part runs through a negotiating channel between finance ministries.

The context the NYT supplies: China comes to these talks with a very uneven hand. AI there is moving forward while the economy is **in its worst shape in decades**. The WSJ, the same day, describes the other side of that coin - last year China registered

more than seven million new one-person companies, and a large share of that entrepreneurship comes out of unemployment and burnout.

Why this matters. When the two largest players sit down to discuss a technology on a dedicated track, it usually means the rules of the game will start being written there, outside the companies. For anyone building products on models, this is an early signal: restrictions on access to hardware, weights and markets arrive from the diplomatic table and change faster than the technology does. The thing to watch is which topics made it onto the working groups' agenda.

TOPIC 3

Big Tech keeps $300bn of AI exposure off balance sheets

SOURCES FT FT

Wall Street has found a way to turn the credit strength of tech giants into cheaper funding for data-centre construction. The mechanism is guarantees: a tech company guarantees the debt of the entity doing the building, and the exposure itself **stays off its balance sheet**. The volume the FT attributes to this structure is around $300bn.

Why this matters. This is the same construction behind old leasing schemes and off-balance-sheet vehicles, and it always means one thing: the risk has not gone anywhere, it has merely stopped being visible in the accounts. For a reader trying to gauge the real scale of the AI build-out, the practical conclusion is that capital expenditure in quarterly reports no longer describes the full picture, and the pace of investment can no longer be judged from those figures alone.

TOPIC 4

Step 5 Preview: 600B parameters, 27B active, 1M context - and Pokemon as a benchmark

SOURCES StepFun Stepfun · discussion Hacker News

China's StepFun showed Step 5 Preview. The architecture is a sparse mixture of experts: **600 billion parameters in total, 27 billion active per token**, a context window of **1 million tokens**, with vision input. On the Artificial Analysis Intelligence Index the model scores **44**. It is promised with open weights.

The most interesting part is how they measure long-horizon autonomy. With no game-specific optimisation the model plays Pokemon and **sustained meaningful progress for more than 3000 turns and 6 million tokens of interaction**. By turn 3082 it had unlocked Cut, earned three Gym Badges and defeated Lt. Surge - roughly a third of the way through the main storyline.

Why this matters. Classic benchmarks measure a single answer, while real agent tasks break on something else: the ability to hold the thread across thousands of steps. A game is

convenient here precisely because progress in it cannot be faked - either the badge is there or it is not. This is a useful way to think about evaluating agents in general: look for a task with a long horizon and an unambiguous external criterion of success, instead of trusting an average score on short examples.

TOPIC 5

Pirate Face: model weights as magnet links that cannot be taken down

SOURCES pirateface.co Pirateface · discussion Hacker News

The service distributes open models - language, image, audio and datasets - as

magnet links instead of downloads from a single host. The idea in one line: a torrent has no owner and no single point of failure, so there is nobody to take the distribution down.

The reaction on HN is striking mostly for how unanimous it is. The refrain: torrents "should really be the preferred method for distributing AI model weights", and this "arguably should've been a thing since day 1".

Why this matters. The question here is broader than piracy. Any team building a product on an open model has to answer what happens if those specific weights vanish from their host tomorrow, whether through licensing, an acquisition or a regulatory demand. The practical conclusion is simple: weights that production depends on are kept in-house rather than left on somebody else's CDN. Torrent distribution is only one way of doing that.

TOPIC 6

Google open-sourced AX, an orchestrator for agent tasks on Kubernetes

SOURCES agentexecutor.io Agentexecutor · discussion Hacker News

AX describes an agent task declaratively, in YAML: a workspace (a git repository with a branch, for instance) and a task with a plain-text goal. The system then sandboxes it, wires up the environment, **fences the network**, and promises scale to billions of tasks per cluster. A task can be suspended and resumed, and files in the workspace survive the pause.

The engineers' reaction is cooler than the technology. The sharpest comment in the thread: "k8sification of AI was always inevitable". Separately, people note that Google has a **parallel project, Scion**, which sits better with existing tooling, whereas AX is more of a greenfield construction of its own.

Why this matters. You can see here what agent infrastructure is turning into: from a library you import into your code, it becomes a cluster scheduler with its own manifests. For teams already running Kubernetes this is a familiar model. But the cost is worth noticing too -

declarative YAML and a bespoke runtime mean one more layer to understand during an incident.

TOPIC 7

Samsung doubles HBM4 output - and bets on glass carriers

SOURCES Seoul Economic Daily Sedaily · discussion Hacker News

Samsung is expected to more than double output of its HBM4 memory family next year.

Capacity is to expand to **250,000 wafers a month**. The sixth and seventh generations (HBM4 and HBM4E) will make up **80% of output**. Separately, the volume of outsourced glass carriers is reported to rise **2.5-fold** - they are needed for stacking large numbers of layers.

Why this matters. Memory bandwidth has long been a tighter constraint on training and inference than the number of compute cores. That makes HBM production plans a more honest indicator of the industry's real pace than model announcements: hardware is planned years ahead and lies poorly. The glass-carrier detail is not a technical trifle here but a hint at where the manufacturing barrier currently sits.

TOPIC 8

Undercover inside a supply-chain hacking gang

SOURCES Ars Technica / WIRED Ars Technica

Google's threat intelligence group said it had **a mole inside the inner circle** of the group known as TeamPCP. This became known after two of its alleged members were arrested and charged in Australia last month.

Why this matters. Supply-chain attacks are the class where an individual company's technical defences decide almost nothing: the target is whoever owns the package you pull in. That is why countering them drifts towards human intelligence and law enforcement. For teams the practical conclusion stays boring and workable: know what is being pulled into the build transitively, and pin dependency versions.

TOPIC 9

"The senior engineer death spiral" - a post holding the top for two days

SOURCES Sunil Pai Sunilpai · discussion Hacker News

The author describes a common failure: an engineer starts a new senior role and begins

play-acting someone more experienced than they are. They take on an oversized design, work 60 to 80 hour weeks to "prove" themselves, and burn out precisely because they tried to close a gap in confidence with volume of work.

The text is written as a monologue to a friend who had just landed a very senior, very well paid role and was asking how to survive that schedule.

Why this matters. It is a rare case of something other than a technology holding the top of HN - a workplace pathology, holding there because people recognise themselves in it. What is useful is that the symptom is named precisely: this is an attempt to close impostor syndrome with volume of work. Telling those two states apart is worth being able to do, both in yourself and in the people you are responsible for.

TOPIC 10

Tao hands the floor to a guest: why do we still need human mathematicians

SOURCES Tao's blog Wordpress · FT FT

A guest post by Po-Shen Loh on Terence Tao's blog. The occasion was a wave of open letters in defence of the mathematical community, which intensified sharply after OpenAI announced a solution to the Millennium Prize variant of Navier-Stokes. The scale of support: the Leiden Declaration has gathered **4000+** signatories, the Math and AI initiative **7000+**.

One detail the author flags himself in a footnote: **100% of the prose was written by a human** in a text editor with no generation, and only the layout and some headings were generated. In a text from 2026 such a footnote is already necessary, which is probably the best illustration of the subject.

The FT, the same day, comes at it from another angle, writing about AI as a powerful but problematic collaborator in mathematics, emphasising what follows when an algorithm reaches a goal under badly defined constraints.

Why this matters. Mathematics here is an early warning for every profession whose value lies in a verifiable result. The question the authors pose is not about replacement but about what remains for a person once an answer becomes cheap to obtain - and their answer is about framing the problem and taking responsibility for the choice, with speed of computation a secondary concern.

---

misc

- **Qwen Image 2.1** - a new version of the image model (Hacker News).

  More interesting than the release itself is an observation in the thread: local image generation currently feels stronger than local code generation, followed immediately by the counterargument that "LLMs are great at what you are not skilled at".
- **Spain has blocked archive.today and its mirrors** Reclaimthenet - by decision of the Second Section of the intellectual property commission, **with no court ruling**. Dated 18.09, it reached the HN front page yesterday.

- **Nobody pays for FOSS** Seldo - a 5000-word longread on the economics of open source via the hawks-and-doves model.

  Published 13.09, the discussion caught fire now.
- **The LLMentalist Effect** Softwarecrisis - **a text from 2023** that fired again: the argument that a chat interface reproduces the mechanics of a psychic's cold reading. The date matters, this is not a fresh publication.
- **Prompts aren't Real** Evaluation - a talk on giving up manual prompt tinkering in favour of building evaluation and optimisation pipelines.
- **RSA-896** Saweis - another size from the old RSA list has been factored.
- **Resident Evil 4 for GameCube** GitHub - a complete byte-identical decompilation to C/C++.

How the topics were chosen

Two layers out of three: Hacker News within the window with a points>80 threshold (39 stories) and the media layer (41 items, 17 of 17 feeds alive). Priority went to anything with a verifiable mechanism or a figure from a primary source.

Deduplication against yesterday's issue was done. Dropped: **Exfiltrate Your Weights** (item 7 yesterday, nothing changed in a day beyond its HN score) and the **"AI Force" announcement** as a standalone story - it survives inside topic 2, now as background to the negotiating track rather than as an event.

Every link was checked with curl before sending.

What I did not verify

🔴 **The X lists were not collected today. This is broken access, a quiet day has nothing to do with it.** Both pinned lists (AI + Product + Tech Radar, Tech + Product)

stayed out of this issue. The cause is specific: the browser profile used for collection **lost its X cookies** - only guest ones remained (`guest_id`, `personalization_id`), with no `auth_token`, `ct0` or `twid`. A live token was found in another profile and moved into place, but the session still would not come up: a headless browser cannot decrypt the cookies without access to the keychain, and that request requires confirmation by hand.

The honest consequence for this issue: today it is a view from the press and Hacker News without the layer of curated subscriptions. First-hand operational news - founders' announcements, researchers' reactions - may have passed by.

Also not verified:

- FT, WSJ and NYT material was read from headlines and feed summaries, full texts behind paywalls. The $300bn figure and the negotiation details come from the publications' summaries rather than the body of the articles.

- Qwen Image 2.1: the release page returns an empty shell without JavaScript, so the description leans on the discussion instead of official specifications.
- Step 5 Preview: the figures are taken from quotes of the official page in the discussion; the site itself serves its content via script.
- The Substack layer was not collected separately today.
